forests

MDPI

Article

Quantifying Risk Levels for Active Safety Systems in Autonomous Forest Machinery Using Vision Language Models

Kengo Usui

Department of Forest Engineering, Forestry and Forest Products Research Institute (FFPRI), Matsunosato, Tsukuba 305-8687, Ibaraki, Japan; usui_kengo790@ffpri.go.jp

Abstract

Forestry is recognized as one of the most dangerous industries in the world. To enhance forestry safety, autonomous machinery and safety systems for such machinery are essential. This study aims to introduce large language models (LLMs)—especially their extensions to images, vision–language models (VLMs)—to enable human-like decision-making for autonomous forest machinery. This research focused on VLMs as an active safety system that can adapt to environments and evaluated the effectiveness of a system that quantitatively makes decisions regarding hazard levels using contrastive language–image pretraining (CLIP). The results of industry type, tree state, and road state classification using pretrained models showed that for three tasks—forestry identification, hung-up tree detection, and road collapse sensing—the target classes consistently exhibited higher similarity with disaster texts compared with nontarget classes. Although the F1 scores were 0.693, 0.324 and 0.634, respectively—indicating that the system is insufficient as a direct active safety system—the application of a similarity threshold optimized to maintain a recall of 0.9 yielded F1 scores of 0.291 and 0.584 for tree state and road state, respectively. These results suggest that the system can potentially be used as a quantitative indicator of hazard by setting a threshold on the similarity score.

Keywords: contrastive language–image pretraining; forestry robot; large language models

1. Introduction

Forestry is a globally recognized high-risk and dangerous industry. The International Labour Organization (ILO) reports that working in hazardous sectors such as agriculture, forestry and fishing, mining, construction, and manufacturing continues to represent the main risk to workers' lives and well-being. Each year, 200,000 fatal injuries occur in these sectors, representing 60% of all fatal occupational injuries [1]. Furthermore, a joint report by the Food and Agriculture Organization, ILO, and the United Nations highlights that work in forests is associated with inherently high occupational safety and health risks, primarily related to the use of heavy machinery, falling trees, climatic hazards, noise, vibration, non-ergonomic working postures, stress and strain, as well as exposure to various chemical and biological substances. Therefore, preventing accidents and reducing health hazards are important features of the research on forest operations and forest engineering [2].

Another challenge in forestry involves the decline in the number of forestry workers in some regions. Lippe et al. [3] reported that for the sub-sector of forestry and logging, visible forest-based employment was approximately 12 million persons full-time equivalent in 2015, showing a slight decrease compared to 2010. In Japan, the number of forestry workers has similarly shown a long-term decreasing trend, with 43,710 people as of fiscal

Check for updates

Academic Editor: Gianni Picchi

Received: 8 May 2026
Revised: 3 June 2026
Accepted: 12 June 2026
Published: 17 June 2026

Copyright: © 2026 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

year 2020 [4]. To address future labor shortages in forestry and ensure worker safety, the mechanization of forestry operations and further development of unmanned and autonomous forestry machinery are urgently necessary. Current automation technology development efforts for forestry machinery focus primarily on harvesting and forwarding operations, where serious occupational accidents are likely to occur because of the handling of heavy objects. For example, automation technologies are being developed for forwarders used in forwarding within cut-to-length systems and grapple loaders used in loading operations [5–8].

The implementation of automated forestry machines in actual forestry operations requires adaptation to various conditions on-site and, most importantly, the assurance of operational safety. To support safety functions compatible with automation, active safety systems should be incorporated with accident-prevention systems.

The implementation of an active safety system requires sensors capable of perceiving the external environment. For example, sensors in the automotive domain include radar, LiDAR, cameras, and ultrasonic sensors [9]. While these sensors are effective in structured outdoor environments such as those encountered by automobiles, forests are not artificially controlled spaces but complex environments consisting of standing trees, understory vegetation, and road structures, among others, that are intermingled and temporally variable. Furthermore, forests in Asia, particularly in Japan, are characterized by steep slopes and abundant microtopography, which create unstable conditions for machine operation. Consequently, the actual conditions in forestry are highly diverse. Accordingly, the effectiveness of individual sensors is substantially limited in such heterogeneous forest environments. For instance, in forested areas, vehicles may need to traverse vegetation; however, although radar and ultrasonic sensors can measure distances, their low point cloud density makes it difficult to obtain reliable object recognition information for assessing the traversability of such vegetation. LiDAR, by contrast, can acquire denser point clouds, and perception pipelines based on LiDAR-driven object detection have become increasingly prevalent [10]. Nevertheless, accurately separating ground surfaces from vegetation in LiDAR point clouds remains challenging [11], making it difficult to determine traversability. In image-based approaches, the applications of image recognition technologies, such as strip road detection using deep learning, have been reported [12,13]. However, methods based on deep learning are highly dependent on training data, and their accuracy may deteriorate in situations different from those used during training. In this context, Zhou et al. [14] pointed out that generalization to out-of-distribution (OOD) data, while natural to humans, remains challenging for machines to reproduce. This is because most learning algorithms strongly rely on the i.i.d. assumption on source/target data, which is often violated in practice due to domain shift. Consequently, this limitation results in numerous forest environments that require advanced decision-making beyond the training data. Therefore, the use of deep-learning-based technologies inherently carries the risk of malfunction due to misdetections, which can directly lead to accidents such as machine overturning, especially during felling and forwarding operations involving large, heavy standing trees and logs. For this reason, active safety systems for forestry should enable more flexible decision-making and integrate multiple factors holistically.

Therefore, this study focuses on large language models (LLMs) as a means of enabling advanced decision-making in autonomous forestry systems. Currently, LLMs refer to transformer language models that contain hundreds of billions (or more) of parameters, which are trained on massive text data [15]. LLMs are effective for various natural language processing problems, such as translation, question-answering, and cloze tasks [16]. Furthermore, active research efforts are directed to extending them to multimodal LLMs

that can handle multiple modalities—especially vision–language models (VLMs) that can process both text and images—to expand their scope of application beyond text.

The industrial application of LLMs is currently vigorous, and their adoption in various industries, such as medicine [17], construction [18], and agriculture [19], is being considered. In forestry, reported applications include expert systems in remote sensing [20] and acquisition of domain knowledge [21]. Meanwhile, VLMs have been applied to forest fire detection [22]. However, despite the extensive application of LLM-based approaches in robotics [23], which use technologies similar to those in forestry automation, their use in autonomous forestry machinery remains limited to a few trials, such as that by Usui [24]. To develop an active safety system that can comprehensively judge multiple risk factors using these technologies, machines must be able to quantify the potential risk of their surrounding environments.

In recent years, contrastive language-image pre-training (CLIP), a VLM that learns relationships between text and images, has enabled the direct evaluation of image–text similarity [25]. This allows for the quantitative evaluation of whether a machine is in an environment similar to past accident cases by comparing an image of its surroundings with text describing hazardous situations. Therefore, this study focuses on VLMs capable of human-like advanced decision-making and aims to evaluate the effectiveness of similarity quantification using CLIP for active safety against diverse environments in forests.

2. Materials and Methods

This study quantifies potential risk by evaluating the similarity between texts describing past accidents and surrounding images using CLIP.

2.1. Evaluation of Forestry Domain Knowledge in Pretrained Models

CLIP performs embedding by converting texts and images into numerical vectors. By calculating the similarity between these images and natural language vector spaces, CLIP can classify broad-range images. Pretrained weights for CLIP and similar models are publicly available, allowing their use without additional training. The training of these models generally utilizes information obtained from web pages, such as LAION-2B/5B [26,27] or DataCompXL [28].

Such training datasets are designed to capture recognition capabilities for general objects rather than specialized domain knowledge, such as that related to forestry. Although applications of CLIP models to forest environments—such as fire detection [29]—have been reported, comparative evaluations of CLIP model performance across different datasets remain scarce, with only a limited number of studies, including Usui [24], addressing this issue. Thus, this study first evaluates the extent to which a CLIP pretrained model has acquired forestry-related domain knowledge.

Furthermore, LLMs/VLMs are primarily designed for English, with a high proportion of English training data. It has been reported that texts generated in languages other than English tend to be less natural than human-authored writing [30]. A decrease in task performance for non-English languages has also been reported [31].

This study computes the similarity between actual accident cases and images, assuming that accident cases are often recorded in the local language of the accident site. To apply VLMs to accident cases described in various languages, two strategies can be considered. The first is to process text in its original language using a corresponding VLM or a multilingual model. The second is to translate the disaster cases into English and then process them using an English VLM. The former enables the preservation of nuances in each language but requires a model for each language or a multilingual model, raising concerns about performance degradation for some languages. Meanwhile, although the

latter risks losing features through translation, it enables the creation of a unified dataset, with English models generally exhibiting higher performance. Therefore, this study adopts the latter approach and conducts evaluations using an English model. For evaluation using local languages and corresponding models, please refer to Usui [24]. Because numerous CLIP models that can handle English and their extensions have been proposed, this study evaluated representative ones, and the models used are shown in Table 1. Hereafter, model names are referred to using their abbreviations presented in Table 1. To focus on architectural and linguistic differences rather than model capacity, we fix the image encoder equivalent to ViT-B across most experiments.

Table 1. Evaluated models and languages.

Model	Abbreviation	Language
open_clip/ViT-B-32/openai [25,32]	CLIP-B32-OpenAI	English
laion/CLIP-ViT-B-32-laion2B-s34B-b79K [25,26,32]	CLIP-B32-Laion	English
open_clip/ViT-B-32/datacomp_xl_s13b_b90k [28,32]	CLIP-B32-Datacomp	English
open_clip/ViT-B-32/metaclip_fullcc [32,33]	CLIP-B32-MetaCLIP	English
BeichenZhang/LongCLIP-B [34]	LONGCLIP-B	English
laion/CLIP-ViT-B-32-xlm-roberta-base-laion5B-s13B-b90k [25,26,32]	CLIP-B32-xlm-roberta	Multilingual
timm/ViT-B-16-SigLIP-i18n-256 [35,36]	SigLIP-B	Multilingual
open_clip/nllb-clip-base/v1 [32,37]	NLLB-CLIP-B	Multilingual
open_clip/nllb-clip-base-siglip/v1 [32,35,37]	NLLB-SigLIP-B	Multilingual

2.2. Evaluation Method

Regarding situations surrounding occupational accidents, Japan's Ministry of Health, Labour and Welfare publicly discloses a portion of data on occupational accidents occurring in Japan [38], from which fatal accident records were used. In this study, "Occupational accident cases" text describing situations in natural language and "Industry types (Medium classification)" indicating sectors such as forestry, retail trade, and construction work were used from 29,737 fatal accident cases between 1999 and 2022. As the accident cases were originally described in Japanese, they were preprocessed through translation to English using gpt-oss:20b via Ollama (version 0.11.11). Examples of the translated accident cases and corresponding industry type are shown in Table 2.

Table 2. Translated occupational accident cases corresponding to the industry type.

Occupational Accident Cases	Industry Types
At a timber harvesting site, a collection winch machine was conducting collection operations. After the collection, the operator removed the load hitching wire and turned the machine to the right in preparation for the next collection. The winch machine tilted toward the valley side, and its bucket struck the victim, who was standing to attach the load hitching wire to the next log, crushing the victim under the bucket.	Forestry
During the renovation of the factory roof, a person affected by the disaster who was working on the slate roof stepped onto the PVC waveboard at the light retrieval point and fell approximately six metres to the floor below.	Building construction
The victim died when his head and right arm were jammed within the hazardous limit (mold) of a 1000 t powered press machine.	Manufacture of transport equipment

The processing flow for this study proceeded as follows. First, all occupational accident cases were vectorized using CLIP. Subsequently, images were similarly vectorized, and the cosine similarity between the image and text vectors was calculated. For each image, the text with the highest cosine similarity was identified as the most similar text. Finally, the predicted label derived from the text was compared with the ground-truth label of the image. The following classification tasks were used as evaluation metrics. First, to assess the acquisition of forestry domain knowledge, the model was evaluated on its ability to classify whether a specified image was related to forestry. In other words, this task measured whether the industry type of the text matched that of the most similar image. Next, to evaluate the system's decision-making capabilities regarding potential risks, two hazardous situations were considered: (1) whether the state during felling constitutes a hung-up tree and (2) road collapse. A hung-up tree is a specific, dangerous situation in forestry that directly evaluates knowledge of the forestry domain. In this study, differences between hung-up trees and the trees they were caught on were not considered; instead, images and texts related to "hung-up" were labeled as "tree state belongs to hung-up class". The second point was road collapse, a dangerous situation that can occur during machinery travel. As similar accident cases exist not only in forestry but also in civil engineering, this evaluates whether the model can correspond to broader, more general knowledge. In this study, situations where forestry machinery cannot pass because of cut slope and fill slope failure or surface erosion were labeled as "road state belongs to road collapse class".

In this study, similarity was evaluated as a quantitative risk metric for the active safety system of forestry machinery. Accordingly, accurate classification of non-hazardous situations is also essential. Therefore, similarity-threshold-based filtering was examined in addition to the non-threshold approach in tree state and road state. After computing cosine similarity, images with values below a predefined threshold were regarded as non-hazardous and classified as belonging to the nontarget classes (not hung-up and not collapsed).

Forestry-class images were collected in Japan between 2006 and 2025 and included forested sites where harvesting and silvicultural operations (e.g., felling and planting) were conducted, as well as spur and forest roads. By contrast, the nonforestry images were randomly sampled from a randomly selected subset of the Google Open Images Dataset [39]. After selecting the required number of images, those captured in forest environments were manually excluded. Consequently, the nonforestry images do not include forest scenes; however, they do include outdoor environments such as trees, pastures, parks, and distant mountain landscapes. The number of images from the Google Open Images Dataset was adjusted so that the counts for forestry and nonforestry classes would be equal. Consequently, 648 images were used for evaluation. Examples of images for each category are shown in Figure 1, and the breakdown of classes within the dataset is shown in Table 3.

Table 3. Number of classes in each category.

	Industry Type		**Tree State**		**Road State**	
	Forestry	**Nonforestry**	**Hung-Up**	**Not Hung-Up**	**Collapsed**	**Not Collapsed**
Num. of images	324	324	33	615	129	519

For annotation, classes were manually assigned to all images. For text annotation, industry types were taken from the "Industry Type (Medium Classification)" field in the accident dataset, resulting in two classes: forestry and nonforestry. Relevant accident cases for tree and road states were isolated for manual classification using two combined filtering

criteria applied to the original Japanese text. Texts were selected if they satisfied either of the following conditions:

- Keyword Presence: The text contained specific Japanese keywords. For tree state, the keyword included "hung-up" (Kakarigi in original); for road conditions, "slope collapse" (Shamen Houkai), "road collapse" (Michi no Houkai), and "cut slope failure" (Norimen Houkai).
- LLM Verification: An LLM (gemma3:4b) verified whether the segment described the same target keywords defined in the Keyword Presence condition. The model was queried using an English prompt template, into which the Japanese keywords and accident cases were inserted. The prompt structure was as follows: "Is the sentence shown below describing [Keyword]? The answer must be Yes or No only. Sentence: [Occupational accident cases]".



(**a**) (**b**)

(**c**) (**d**)

Figure 1. Examples of images: (**a**) industry type: forestry; (**b**) tree state: hung-up; (**c**) road state: collapsed; and (**d**) industry type: nonforestry.

Following this screening process, the filtered texts were manually reviewed, and the final classification labels were assigned. For these classification tasks, accuracy, precision, recall, and F1 score were used as evaluation metrics. Because the class distributions for the tree state and road state categories were imbalanced, the F1 score—being robust to such imbalance—was selected as the primary evaluation metric. All processing was performed on a single computer (CPU Intel Core i9-10900K @ 3.70GHz, Memory 96GB, GPU NVIDIA GeForce RTX 5090, OS Ubuntu 24.04).

3. Results

3.1. Image–Text Similarity by Class and Model

First, the mean and standard deviation of image–text similarity for each category and model are shown in Table 4.

Table 4. Image–text similarity for each category.

	Industry Type		Tree State		Road State	
	Forestry	**Nonforestry**	**Hung-Up**	**Not Hung-Up**	**Collapsed**	**Not Collapsed**
CLIP-B32-OpenAI	0.316 ± 0.016	0.264 ± 0.022	0.319 ± 0.014	0.288 ± 0.033	0.318 ± 0.016	0.283 ± 0.032
CLIP-B32-Laion	0.351 ± 0.024	0.260 ± 0.034	0.352 ± 0.024	0.303 ± 0.054	0.360 ± 0.020	0.292 ± 0.051
CLIP-B32-Datacomp	0.299 ± 0.020	0.219 ± 0.030	0.298 ± 0.022	0.257 ± 0.048	0.305 ± 0.017	0.248 ± 0.046
CLIP-B32-MetaCLIP	0.330 ± 0.018	0.275 ± 0.020	0.333 ± 0.022	0.301 ± 0.033	0.336 ± 0.016	0.295 ± 0.032
LONGCLIP-B	0.327 ± 0.010	0.276 ± 0.015	0.330 ± 0.010	0.300 ± 0.029	0.330 ± 0.010	0.294 ± 0.028
CLIP-B32-xlm-roberta	0.281 ± 0.023	0.195 ± 0.033	0.285 ± 0.020	0.236 ± 0.051	0.282 ± 0.017	0.227 ± 0.051
SigLIP-B	0.140 ± 0.018	0.079 ± 0.028	0.155 ± 0.012	0.107 ± 0.037	0.142 ± 0.015	0.102 ± 0.038
NLLB-CLIP-B	0.361 ± 0.042	0.224 ± 0.057	0.354 ± 0.036	0.289 ± 0.086	0.361 ± 0.040	0.276 ± 0.085
NLLB-SigLIP-B	0.160 ± 0.025	0.061 ± 0.040	0.173 ± 0.021	0.107 ± 0.059	0.160 ± 0.021	0.098 ± 0.060

The mean similarity values differed depending on the model. For all categories—industry type, tree state, and road state—the images belonging to the forestry, hung-up, and collapsed classes tended to show higher similarity with accident cases relative to images belonging to the nonforestry, not hung-up, and not collapsed classes. The Shapiro–Wilk test and the Brunner–Munzel test were performed to compare the target classes (forestry, hung-up, and collapsed) and non-target classes (nonforestry, not hung-up, and not collapsed), and statistically significant differences ($p < 0.01$) were observed across all categories for all models.

In terms of similarity, across all classification tasks—industry type, tree state, and road state—images classified as forestry, hung-up, and collapsed showed higher similarity with texts indicating fatal occupational accident cases than images not in these respective classes. For example, in the tree state of CLIP-B32-OpenAI, the mean similarity for the hung-up class is 0.319, compared with 0.288 for the not hung-up class. Although the mean similarity values differed by model, consistently high similarity was obtained for the target classes from the accident cases. For example, when the industry type is forestry, the mean similarity values vary substantially across models, ranging from 0.140 in SigLIP-B to 0.361 in NLLB-CLIP-B. Nevertheless, in all models, the mean similarity for the forestry class is consistently higher, falling within the range of 0.05–0.14.

To employ the proposed system as an active safety system for automated forestry machinery, real-time operation is required. Table 5 presents the per-image inference time of the models used in this study. Because model loading and text tokenization were performed in advance, the measured inference time corresponds to the duration from image input to the extraction of the most similar text. The processing time per image ranged from 0.0784 s (LONGCLIP-B) to 0.2160 s (NLLB-SigLIP-B).

Table 5. Processing time of each model.

	Processing Time per Image (s)
CLIP-B32-OpenAI	0.0991
CLIP-B32-Laion	0.0980
CLIP-B32-Datacomp	0.0997
CLIP-B32-MetaCLIP	0.0995
LONGCLIP-B	0.0784
CLIP-B32-xlm-roberta	0.0973
SigLIP-B	0.1371
NLLB-CLIP-B	0.0977
NLLB-SigLIP-B	0.2160

3.2. Evaluation of Estimation Accuracy

The estimation accuracies for all categories are shown in Table 6. Additionally, examples of texts with the highest similarity to an image in CLIP-B32-xlm-roberta are presented for cases of correct classification (Figure 2) and incorrect classification (Figure 3).



During a felling operation focused on sugi, at a site with a slope gradient of 43.5 degrees, mizuki was felled toward the valley side and became the underlay of the felled tree.

(**a**)



While inspecting the highway in shifts after it had been closed due to the approach of a typhoon, we discovered a bulge on the cut slope of the downhill lane. When we approached the site, the slope collapsed on a large scale, sweeping four people to the guardrail of the central median, and burying two people in landslide material.

(**b**)



Because a minke whale about 5.6 m in length had entered the large gill net set offshore, during the whale hauling operation an attempt was made to haul the whale from the net by tying a rope to its tail and using a roller (winch). The whale flailed, and its tail fin struck the head of a nearby disaster victim directly.

(**c**)

Figure 2. Correct pairs of images and texts: (**a**) industry type: forestry; (**b**) road state: collapsed; and (**c**) industry type: nonforestry.

Table 6. Accuracy of each category.

	Industry Type				Tree State				Road State			
	Accuracy	**Precision**	**Recall**	**F1 Score**	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**Accuracy**	**Precision**	**Recall**	**F1 Score**
CLIP-B32-OpenAI	0.694	0.914	0.429	0.584	0.934	0.250	0.152	0.189	0.823	0.667	0.217	0.327
CLIP-B32-Laion	0.667	0.929	0.361	0.520	0.941	0.000	0.000	0.000	0.826	0.711	0.209	0.323
CLIP-B32-Datacomp	0.662	0.913	0.358	0.514	0.938	0.182	0.061	0.091	0.826	0.750	0.186	0.298
CLIP-B32-MetaCLIP	0.667	0.915	0.367	0.524	0.935	0.200	0.091	0.125	0.858	0.930	0.310	0.465
LONGCLIP-B	0.724	0.901	0.503	0.646	0.935	0.000	0.000	0.000	**0.884**	0.855	**0.504**	**0.634**
CLIP-B32-xlm-roberta	**0.756**	**0.937**	**0.549**	**0.693**	**0.952**	**0.600**	0.182	0.279	0.827	**1.000**	0.132	0.233
SigLIP-B	0.648	0.929	0.321	0.477	0.940	0.000	0.000	0.000	0.812	0.621	0.140	0.228
NLLB-CLIP-B	0.671	0.854	0.414	0.557	0.923	0.293	**0.364**	**0.324**	0.807	0.583	0.109	0.183
NLLB-SigLIP-B	0.687	0.824	0.475	0.603	0.929	0.276	0.242	0.258	0.819	0.773	0.132	0.225

The bold indices denote the models achieving the highest values within this dataset, while underlined indices indicate the second-highest values.

In terms of estimation accuracy, CLIP-B32-xlm-roberta showed the highest values for all metrics for industry type. For both tree state and road state, the model achieving the highest performance varied depending on the evaluation metric. For tree state, CLIP-B32-xlm-roberta achieved the highest accuracy and precision, whereas NLLB-CLIP-B achieved the highest recall and F1 score. For road state, LONGCLIP-B achieved the highest accuracy, recall, and F1 score, while CLIP-B32-xlm-roberta achieved the highest precision. The best F1 scores for each category were 0.693, 0.324 and 0.634, respectively.



In the forest road construction work, while dropping the obstruction trees that had been pre-cut and temporarily positioned on the slope down to the lower part of the slope, we were handling the cut trees one by one from the uppermost ones. One of the trees became snagged in the middle of the slope, and while we were descending the slope and turning the cut tree with a skidder, several of the cut trees above the slope collapsed and fell onto the lower side, landing as a pile.

(**a**)



After the water collection pit work at the mine expansion site was finished, a one-box van was descending the Nishiyama hiking trail from the work yard on the upper side of the mine toward the on-site office at the lower side of the mine. While crossing the runaway prevention embankment that had been constructed 2.8 km from the work yard along the curve of the trail, the van overturned approximately 50 m down the trail. The van carried three occupants; the driver sustained serious injuries from a cervical spine fracture, and the two occupants in the rear seat were killed.

(**b**)

Figure 3. *Cont.*



During the weeding operation, I was stung by a wasp that had nested in the grass beneath a cedar tree.

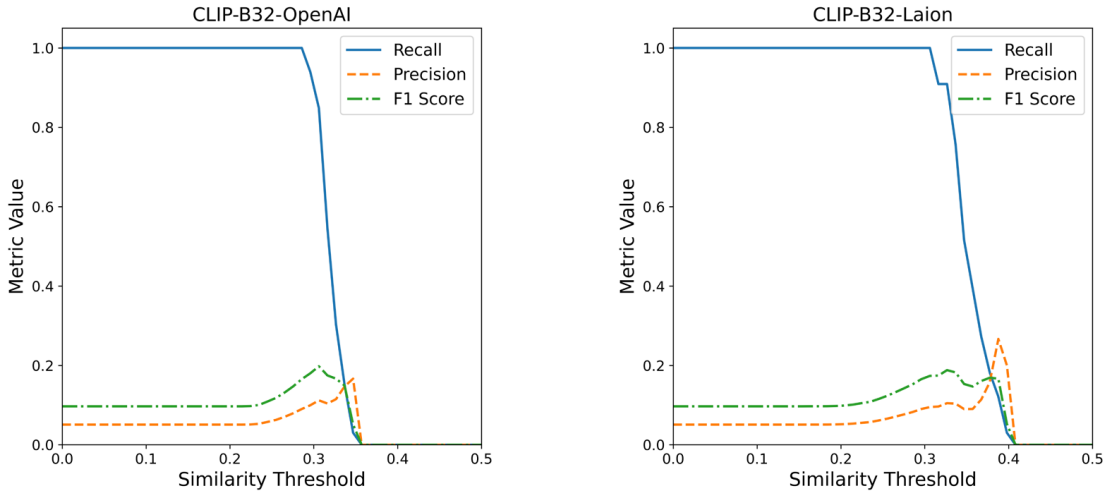
(**c**)

Figure 3. Incorrect pairs of images and texts: (**a**) industry type: forestry; (**b**) road state: collapsed; and (**c**) industry type: nonforestry.

Figures 4 and 5 show the variation in precision, recall, and F1 score as the similarity threshold is adjusted. Because minimizing missed risks is critical in this study, target recall levels were defined, and the corresponding similarity thresholds were determined to yield recall values as close as possible to each target. Target recall values were set to 0.7, 0.8, and 0.9, corresponding to false negatives of 0.3, 0.2, and 0.1, respectively. The resulting target recall levels, similarity thresholds, and performance metrics are summarized in Tables 7 and 8.

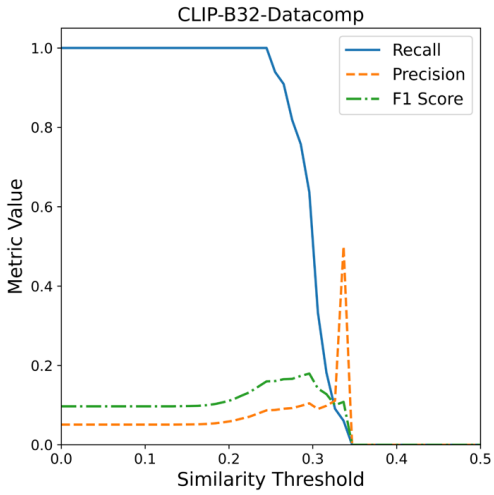
Table 7. Thresholds and corresponding metrics for target recall rates in tree state.

	Target Recall	Recall on the Nearest Target Recall	Similarity Threshold	Precision	F1 Score
CLIP-B32-OpenAI	0.7	0.727	0.313	0.116	0.200
	0.8	0.818	0.309	0.116	0.203
	0.9	0.909	0.301	0.105	0.189
CLIP-B32-Laion	0.7	0.727	0.338	0.101	0.177
	0.8	0.818	0.331	0.101	0.180
	0.9	0.909	0.327	0.105	0.188
CLIP-B32-Datacomp	0.7	0.727	0.289	0.103	0.180
	0.8	0.818	0.279	0.097	0.173
	0.9	0.909	0.269	0.093	0.169
CLIP-B32-MetaCLIP	0.7	0.727	0.318	0.097	0.171
	0.8	0.818	0.314	0.097	0.173
	0.9	0.909	0.302	0.091	0.165
LONGCLIP-B	0.7	0.727	0.326	0.128	0.217
	0.8	0.818	0.324	0.131	0.226
	0.9	0.909	0.316	0.108	0.194
CLIP-B32-xlm-roberta	0.7	0.727	0.273	0.114	0.198
	0.8	0.909	0.265	0.114	0.203
	0.9	0.909	0.265	0.114	0.203
SigLIP-B	0.7	0.758	0.147	0.203	0.321
	0.8	0.818	0.143	0.185	0.302
	0.9	0.909	0.138	0.173	0.291
NLLB-CLIP-B	0.7	0.727	0.343	0.107	0.186
	0.8	0.818	0.326	0.095	0.171
	0.9	0.909	0.296	0.088	0.161
NLLB-SigLIP-B	0.7	0.758	0.166	0.176	0.286
	0.8	0.818	0.165	0.178	0.292
	0.9	0.909	0.145	0.130	0.227

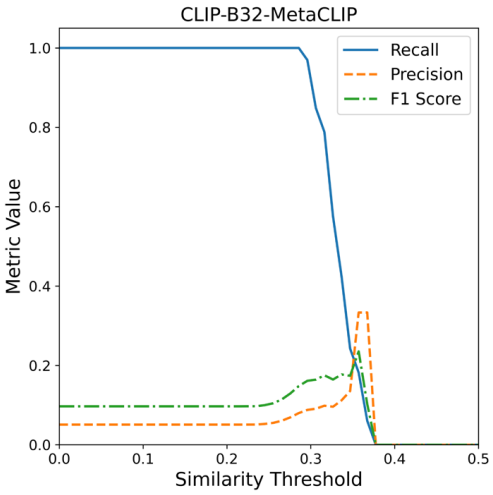


(**a**)

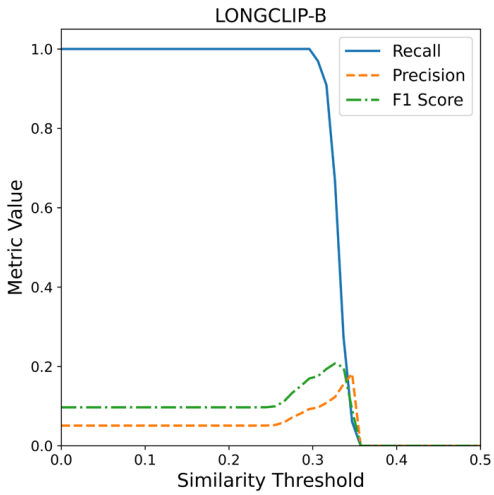
(**b**)



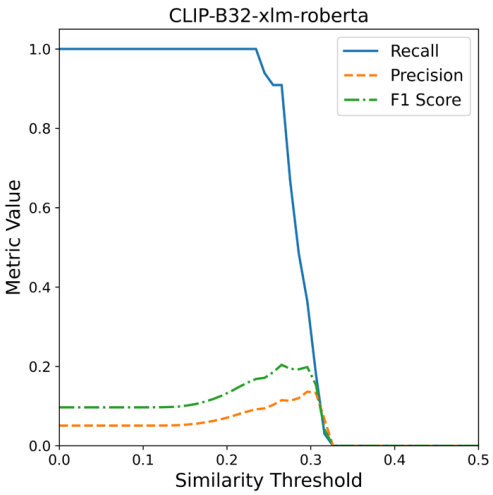
(**c**)



(**d**)



(**e**)



(**f**)

Figure 4. *Cont.*

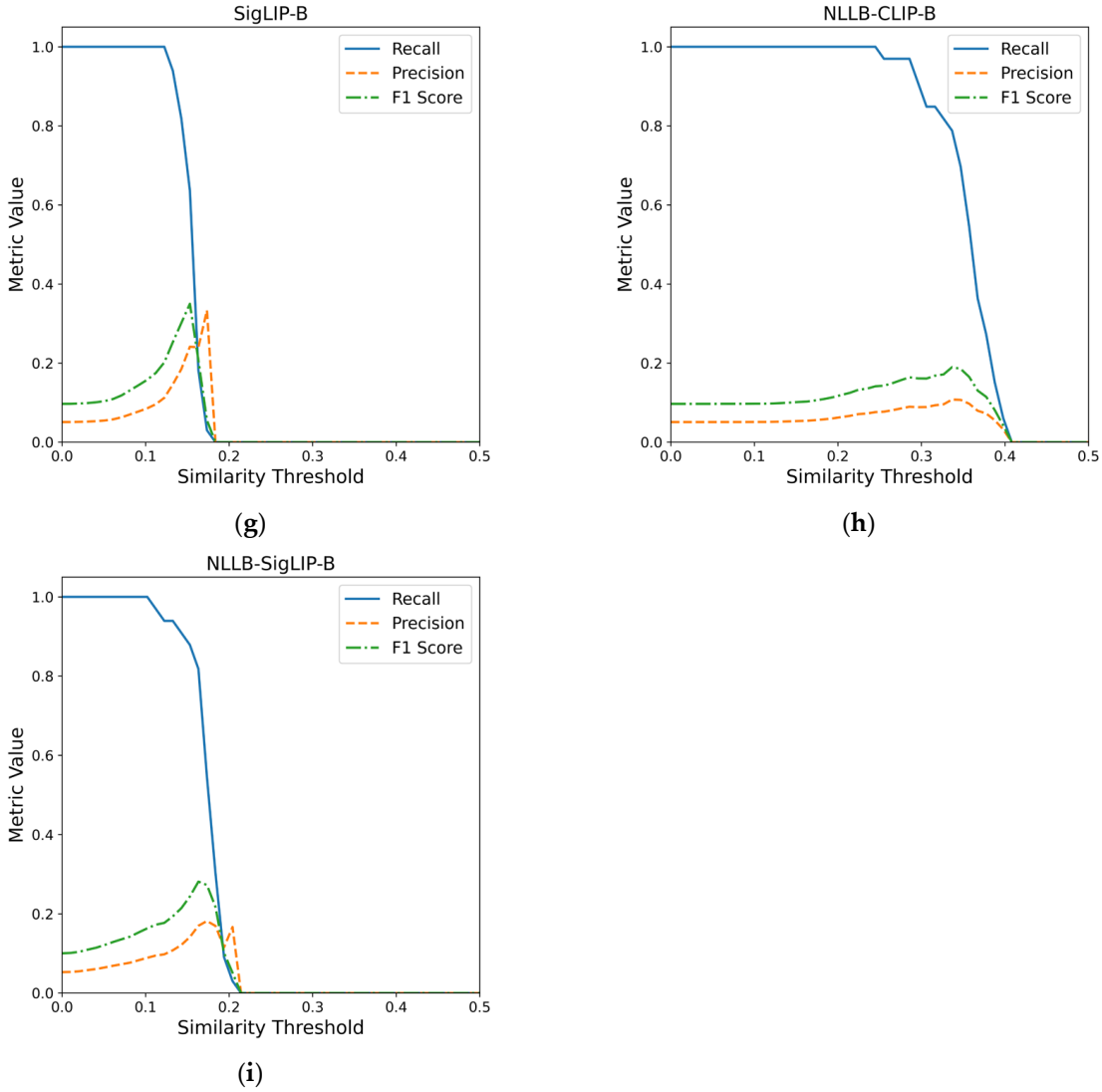


Figure 4. Evaluation metrics across varying similarity thresholds in tree state: (**a**) CLIP-B32-OpenAI; (**b**) CLIP-B32-Laion; (**c**) CLIP-B32-Datacomp; (**d**) CLIP-B32-MetaCLIP; (**e**) LONGCLIP-B; (**f**) CLIP-B32-xlm-roberta; (**g**) SigLIP-B; (**h**) NLLB-CLIP-B; and (**i**) NLLB-SigLIP-B.

Table 8. Thresholds and corresponding metrics for target recall rates in road state.

	Target Recall	Recall on the Nearest Target Recall	Similarity Threshold	Precision	F1 Score
CLIP-B32-OpenAI	0.7	0.705	0.310	0.404	0.514
	0.8	0.822	0.303	0.397	0.535
	0.9	0.922	0.297	0.386	0.545
CLIP-B32-Laion	0.7	0.713	0.353	0.554	0.624
	0.8	0.806	0.343	0.486	0.606
	0.9	0.938	0.333	0.465	0.622
CLIP-B32-Datacomp	0.7	0.705	0.298	0.474	0.567
	0.8	0.814	0.290	0.459	0.587
	0.9	0.907	0.279	0.419	0.574
CLIP-B32-MetaCLIP	0.7	0.705	0.327	0.462	0.558
	0.8	0.806	0.321	0.450	0.578
	0.9	0.915	0.313	0.418	0.574

Table 8. *Cont.*

	Target Recall	Recall on the Nearest Target Recall	Similarity Threshold	Precision	F1 Score
LONGCLIP-B	0.7	0.713	0.324	0.447	0.549
	0.8	0.806	0.321	0.441	0.570
	0.9	0.915	0.317	0.429	0.584
CLIP-B32-xlm-roberta	0.7	0.721	0.273	0.443	0.549
	0.8	0.814	0.269	0.441	0.572
	0.9	0.907	0.258	0.418	0.572
SigLIP-B	0.7	0.705	0.133	0.429	0.534
	0.8	0.806	0.129	0.411	0.545
	0.9	0.907	0.124	0.411	0.565
NLLB-CLIP-B	0.7	0.705	0.339	0.378	0.492
	0.8	0.814	0.324	0.365	0.504
	0.9	0.915	0.301	0.358	0.514
NLLB-SigLIP-B	0.7	0.713	0.147	0.407	0.518
	0.8	0.822	0.140	0.411	0.548
	0.9	0.907	0.134	0.413	0.568

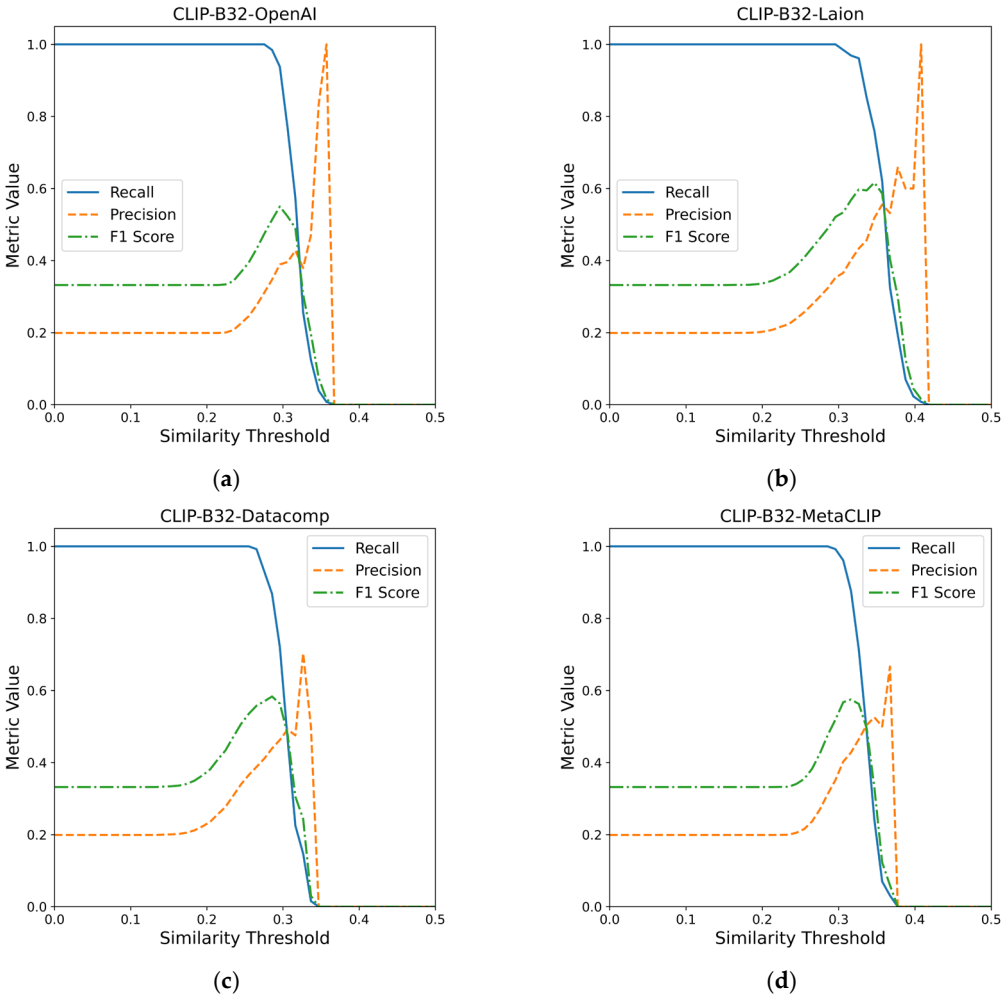


Figure 5. *Cont.*

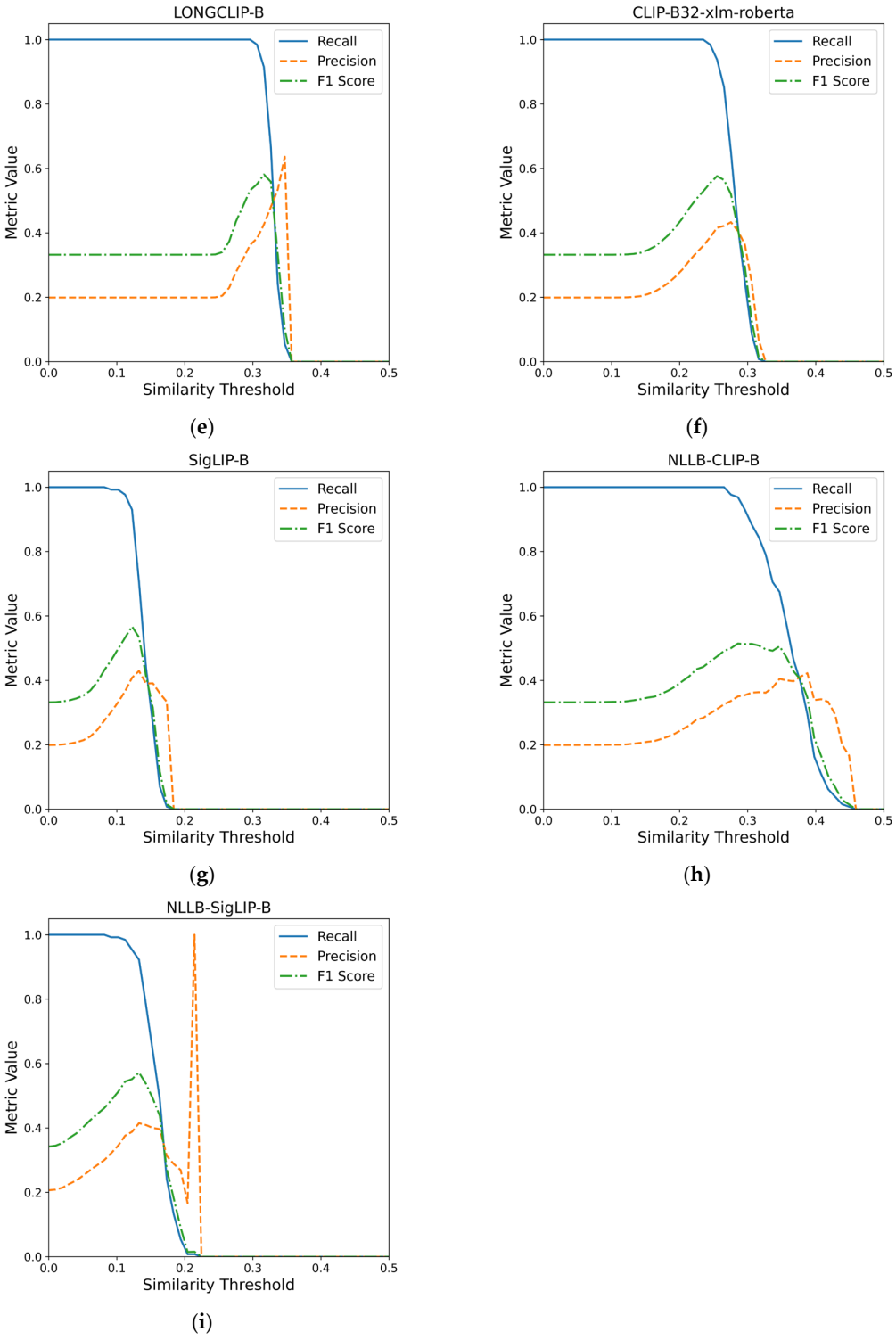


Figure 5. Evaluation metrics across varying similarity thresholds in road state: (**a**) CLIP-B32-OpenAI; (**b**) CLIP-B32-Laion; (**c**) CLIP-B32-Datacomp; (**d**) CLIP-B32-MetaCLIP; (**e**) LONGCLIP-B; (**f**) CLIP-B32-xlm-roberta; (**g**) SigLIP-B; (**h**) NLLB-CLIP-B; and (**i**) NLLB-SigLIP-B.

4. Discussion

4.1. Image–Text Similarity by Category and Model

As shown in Section 3.1, the target class consistently exhibited higher similarity than the non-target class across all classification tasks and models. This indicates that the situations depicted within the images were similar to those during fatal accidents. In other words, images belonging to the forestry, hung-up, and collapsed classes were considered

closer to the circumstances of fatal accidents compared with the nontarget images. That is, the similarity with occupational accident cases was quantified to some extent. This suggests that the similarity between the images and texts describing occupational accident cases could potentially be used as a quantitative indicator of potential environmental risk.

4.2. Estimation Accuracy

In this study, the accuracies for industry type, tree state, and road state were 0.648–0.756, 0.923–0.952, and 0.807–0.882; the precision values were 0.824–0.937, 0.000–0.293, and 0.583–1.000; and the F1 scores were 0.477–0.693, 0.000–0.324, and 0.183–0.634, respectively. As noted in the Materials and Methods section, given the class imbalance, the F1 score, which represents both precision and recall, was used as the primary metric for discussion. General classification accuracies for CLIP and its derivative models were reported as 66.56% for CLIP-B32-Laion on ImageNet-1k, 69.17% for CLIP-B32-DataComp, 24.32% for NLLB-CLIP-B, and 62.36% for CLIP-B32-xlm-roberta [32].

Relative to classification results for general datasets, the accuracies for the target classifications in this study showed comparable results. However, these values are clearly insufficient as detection performance for an active safety system. In contrast, when tree state and road state classifications were evaluated using similarity thresholds, the F1 scores reached 0.291 (SigLIP-B for tree state) and 0.584 (LONGCLIP-B for road state), even at a target recall of 0.9, corresponding to an allowable false negative of 10%. These results suggest that by appropriately setting the similarity threshold, hazardous situations can be detected with minimal risk of omission. Furthermore, as shown in Figures 4 and 5, increasing the similarity threshold generally leads to a decrease in recall and an increase in precision across all the models. As a result, the F1 score exhibits a peak at a specific threshold value. Notably, recall decreases sharply as the threshold increases; even when the target recall is relaxed from 0.9 to 0.7, the corresponding threshold range remains narrow (0.008–0.047 for tree state and 0.007–0.038 for road state). This indicates that model performance is highly sensitive to threshold selection, underscoring the importance of careful threshold tuning for each model. Although the F1 score for tree state is low, the system may still be acceptable from a safety perspective, as a high recall of 0.9—ensuring minimal missed detections—is maintained.

When comparing English-only and multilingual models, differing trends were observed: road state favored English-only models. Meanwhile, no consistent trend was seen for industry type and tree state. Furthermore, because these differences were particularly high in the top-performing and second-best-performing models, they were attributed to model performance rather than language comparison.

Regarding input language, Usui [24] reported classification accuracy when Japanese data were input into a CLIP model in forestry. For industry type, tree state, and road state, the accuracy of the multilingual model "laion/CLIP-ViT-B-32-xlm-roberta-base-laion5B-s13B-b90k" [25,32] achieved F1 scores of 66.83%, 50.79%, and 70.88%, respectively, whereas the scores were 57.62%–64.48%, 48.33%–53.25%, and 48.65%–65.31% for a Japanese model. In this study, for the same model, "CLIP-B32-xlm-roberta," the F1 scores were 0.693, 0.279, and 0.233 for industry type, tree state, and road state, respectively. It is well-established that LLMs achieve the highest performance when the input language is English [40]. The results of this study, together with those of Usui [24], indicate that while these models are effective for general classification tasks, such as distinguishing between forestry and nonforestry data, their performance is limited for domain-specific tasks, including road collapse detection and identification of hung-up trees. When regional variability in forest environments and forestry operations is substantial, even translation into English may not mitigate the wide distribution of region-specific cases within the dataset, potentially

limiting the model's ability to generalize effectively. Similar phenomena have been reported in tasks involving specific non-English-speaking cultural contexts, where performance tends to degrade when using English-centric datasets [41]. These observations suggest that insufficient generalization may arise for underrepresented data in such datasets. Consequently, training with data tailored to the target region and operational conditions may enable more accurate detection.

One possible explanation for the observed differences in accuracy across all models is the limited representation of forestry-related data in the training datasets. In this study, pretrained models with similar architectures but trained on different datasets were compared: CLIP-B32-OpenAI was trained on OpenAI's proprietary dataset, CLIP-B32-Laion on Laion2B, CLIP-B32-Datacomp on DataComp-XL, and CLIP-B32-MetaCLIP on MetaCLIP. For industry type, tree state, and road state, the F1 scores were 0.514–0.584, 0.000–0.189, and 0.298–0.465, respectively, with CLIP-B32-OpenAI showing the highest accuracy for industry type and tree state and CLIP-B32-MetaCLIP for road state. Among the three classifications evaluated in this study, no single model consistently showed high accuracy, making it difficult to identify a dataset capable of acquiring forestry domain knowledge.

Although variations in accuracy attributable to the training datasets were observed among models with the same architectures, overall performance was more strongly influenced by the model itself than by the dataset alone. Because the classification accuracy in forestry is comparable to or higher than that of general datasets, general web-crawled datasets may be sufficient for basic forestry applications. However, for more detailed adaptation to forestry and application of specialized terminology, continuous domain-specific adaptation is indispensable. For LLMs, Tan et al. [21] proposed a model structure assuming domain knowledge updates aimed at acquiring forestry domain knowledge, suggesting that such continuous learning is necessary. For VLMs, although datasets like AgriClip [19] exist for agriculture and fisheries, public datasets specific to forestry are limited to applications like image segmentation, with no dedicated VLM dataset. Thus, expanding multimodal forestry-specific datasets is essential for learning forestry domain knowledge efficiently.

Another important factor is the relatively short context length of standard CLIP models. For example, OpenAI's original implementation of CLIP [25] limited the context length to 77 tokens. The average length of occupational accident cases in this study was 308 characters (equivalent to ~77 tokens for English input), with a maximum of 923 characters, exceeding this limit. Consequently, information from these accident cases could potentially be lost by some models. In contrast, LONGCLIP-B, which supports longer context lengths, recorded the second-highest F1 score for industry type and the highest for road state. This suggests that for classifications using relatively long text descriptions, such as those of occupational accident cases, higher accuracy classification is possible by using models with a longer supported context length.

4.3. Model Limitations

The models evaluated in this study were intended for implementation as active safety systems on forestry machinery. Therefore, we focused on the tendencies during misclassification, particularly on image–text pairs that resulted in false negatives, and examined all images that were incorrectly classified by any model. An example of an image incorrectly classified across all models for industry type is shown in Figure 6.

For industry classification, the occupational accident cases most similar to the image, identified by the top two models, are as follows:

CLIP-B32-xlm-roberta: After the water collection pit work at the mine expansion site was finished, a one-box van was descending the Nishiyama hiking trail from the work yard

on the upper side of the mine toward the on-site office at the lower side of the mine. While crossing the runaway prevention embankment that had been constructed 2.8 km from the work yard along the curve of the trail, the van overturned approximately 50 m down the trail. The van carried three occupants; the driver sustained serious injuries from a cervical spine fracture, and the two occupants in the rear seat were killed.



Figure 6. Incorrect image for industry type.

LONGCLIP-B: During slope conservation work for slope maintenance, the working road became impassable because surface soil displaced by a mini excavator had blocked it; when we attempted to pass the upper side and swing around to the opposite side, we fell approximately 40 m down the slope.

An example of an image incorrectly classified for tree state is shown in Figure 7.



Figure 7. Incorrect image of hung-up trees.

For tree state, the occupational accident cases most similar to the image, identified by the top two models, are as follows:

NLLB-CLIP-B: During the felling operation of a standing tree, the standing tree served as the underlay for the branches of the felled tree.

CLIP-B32-xlm-roberta: At a clear-cut site of cedar (sugi and hinoki), we were unable to contact the disaster victim who had been engaged in felling operations. When we went to the site, we found the victim lying prone beside the felling root (cedar A); a fallen cedar B (height 12 m, diameter at breast height 10 cm) had overturned and was covering the victim. Immediate emergency measures were taken, but death was confirmed several hours later.

For slope collapse, an example of an incorrect image is shown in Figure 8.



Figure 8. Incorrect image of slope collapse. Image content was blurred for privacy.

For road state, the occupational accident cases most similar to the image, identified by the top two models, are as follows:

LONGCLIP-B: A disaster victim was traveling in a vehicle on a temporarily established working road within the site when he or she fell into a scrub forest beneath a cliff and was trapped between the vehicle and the ground. The fall point was about ten metres from the road.

CLIP-B32-MetaCLIP: While transporting the soil and sediment generated at the site with a 2-ton dump truck to the residual earth disposal area, the vehicle fell from the roadside onto the cliff face below (approximately 70 m). The site has a gentle curve and no guardrails are installed.

For industry type, a clear similarity was observed between the images and cases. The case from CLIP-B32-xlm-roberta belongs to mining, whereas that of LONGCLIP-B belongs to civil engineering. Both cases described similar circumstances—specifically, a road with slopes and vehicle involvement—with LONGCLIP-B explicitly describing an excavator. These industries perform comparable tasks, such as road development and maintenance, suggesting that the models appropriately selected fundamentally similar operations. Regarding the road state, Figure 8 shows white vehicles on-site, leading to

the identification of overturning or falling accidents as the most similar cases. The tree state reveals clear patterns: both cases involve felling operations. Notably, NLLB-CLIP-B identified a "hung-up tree" case despite the accident case not explicitly using the term "hung-up". This discrepancy likely stemmed from the annotation process, where task-relevant keywords were prefiltered before manual labeling. Consequently, accident cases describing equivalent cases without these designated terms were excluded from their respective categories. In practice, however, high similarity with the image persists even in the absence of such keywords, indicating that similarity evaluations remain sufficiently accurate. In this study, text labeling was performed using a combination of keyword-based search and LLM-driven retrieval. However, the LLM employed in this study resulted in some missed label assignments. This limitation could potentially be addressed by using more advanced LLMs, which may better capture cases that do not explicitly contain predefined keywords.

A common feature among these three classification examples is that, despite the incorrect category classification, the content of the text corresponds to the image, indicating appropriate calculation of similarity. Although this study evaluated models from a classification-based perspective, these VLMs can appropriately evaluate image–text similarity when used as preventive safety systems for forestry machinery.

Meanwhile, images containing distinctive objects elicited strong reactions to those specific objects, resulting in relatively low similarity scores for the other elements. For example, when an image depicted a collapsed road and vehicles, the model showed higher similarity with vehicle-related accident cases while assigning relatively lower similarity evaluations to cases concerning road collapses. This is attributable to the model architecture of CLIP, which calculates similarity indices for entire sentences without placing weights on specific words or terms. Therefore, with the current model structure, performing hazard judgments is difficult for specific dangerous situations in forestry, such as hung-ups or machine imbalance and off-balance. Accordingly, VLMs that perform weighting on specific keywords, such as SToRi [42], have been proposed. The use of these models or the calculation of vectors for each word and weighted analysis in the similarity space enables judgments of similarity or hazard levels for specific keywords. On the other hand, prompt-level methods that perform weighted analysis in the similarity space are also effective. The latter approach involves computing vector representations for individual keywords and weighting their similarity to the input image, enabling the assessment of hazard relevance. Specifically, similarity is computed between the image and keywords representing forestry-specific hazard factors, and these scores are integrated with the overall accident scenario similarity. This framework enables the retrieval of accident cases that are highly relevant in terms of hazard characteristics, even in the absence of direct visual correspondence with primary objects.

4.4. Practical Feasibility

Regarding the practical feasibility of the proposed system, real-time operation is required. As shown in Table 5, most of the models showed a processing time of less than 0.01 s. These results indicate that all models are sufficiently fast for use in sensing systems and capable of real-time operation.

The occupational accident case database used in this study is limited to a maximum length of 1000 characters and therefore consists of relatively short texts. However, when considering the input of more detailed accident reports spanning several pages, the CLIP model employed in this study may not perform as expected, as it is primarily designed for short-context texts such as image captions. To address this limitation, one potential approach is to compress the accident-related information in advance using other LLMs.

Furthermore, Zhou et al. [43] have shown that the performance of CLIP can be improved through prompt engineering, suggesting that optimizing the text inputs may further improve accuracy.

This study suggested vision-based risk metric for the active safety system. However, its performance may be degraded or impaired under certain environmental conditions, potentially limiting the effectiveness of vision-based safety systems in forestry machinery. Weather conditions, particularly reduced visibility due to fog or rain, can continuously affect system performance. In addition, rapid illumination changes caused by trees, as well as transient occlusions from obstacles such as branches, may intermittently disrupt the functionality of vision-based safety systems. In situations where sensor performance is persistently compromised, such as during severe weather, suspending the operation of autonomous machinery is a reasonable safety measure. However, transient failures of vision-based systems are inherently difficult to predict. To mitigate these risks, a multi-layered safety strategy is required, including the integration of complementary sensors such as FMCW LiDAR, which is more robust to adverse weather conditions than conventional LiDAR [9], as well as the dynamic weighting of sensor reliability in response to changing environmental conditions.

In this study, we evaluated a VLM-based quantitative hazard judgment system as an element of technology for active safety in material production sites. Although this study showed that CLIP-based similarity can be used as a quantitative metric, it also suggested that methodological improvements may be necessary for estimating causal objects. ForestGPT [44], an LLM specialized for forest management, has also been proposed, and its use is expected to contribute to ensuring safety in forest areas.

5. Conclusions

This study focused on LLMs/VLMs as active safety systems that can flexibly adapt to environments in autonomous forestry machinery and evaluated the effectiveness of a system that quantifies potential environmental risks using a VLM. The results of industry type, tree state, and road state classification using pretrained models showed that for three classifications—forestry, hung-up, and collapsed—the target classes consistently showed higher similarity with fatal occupational accident cases compared with the nontarget classes. Although the F1 scores were 0.693, 0.324 and 0.634, respectively—indicating the system's insufficiency as a direct active safety system—the application of a similarity threshold optimized to maintain a recall of 0.9 yielded F1 scores of 0.291 and 0.584 for tree state and road state, respectively. These results suggest that the system can potentially be used as a quantitative indicator of hazards by setting a threshold on the similarity score. Although the classification accuracies were 0.694, 0.324, and 0.634, respectively—indicating that the system is insufficient for direct use as an active safety system—the application of a similarity threshold optimized to maintain a recall of 0.9 yielded F1 scores of 0.291 and 0.584 for tree state and road state, respectively. These results suggest that the system can potentially be used as a quantitative indicator of hazard by setting a threshold on similarity.

Funding: This work was supported by JSPS KAKENHI, Grant Number 25K18268.

Data Availability Statement: The occupational accident data presented in this study are available in Shokuba No Anzen Site (Workplace Safety Website) at https://anzeninfo.mhlw.go.jp/user/anzen/tok/anst00.html, accessed on 2 February 2026. A portion of the image data is available in the Google Open Images Dataset at https://g.co/dataset/openimages/, accessed on 8 May 2025. The remaining image data in this article are not readily available because they include image data subject to privacy considerations and are part of an ongoing study.

Acknowledgments: We are deeply grateful to Takumi Uemura and Hidenori Suzuki of the FFPRI for providing image data and their valuable advice. During the preparation of this study, the author used CLIP and its variants listed in Table 1 for the purposes of image analysis. The author has reviewed and edited the output and takes full responsibility for the content of this publication. During the preparation of this study, the author used CLIP and its variants listed in Table 1 for the purposes of image analysis. The author has reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. International Labour Organization. *A Call for Safer and Healthier Working Environments*; ILO: Geneva, Switzerland, 2023. [CrossRef]
2. FAO; ILO; United Nations. *Occupational Safety and Health in the Future of Forestry Work*; FAO: Rome, Italy; ILO: Geneva, Switzerland; UNECE (United Nations Economic Commission for Europe): Geneva, Switzerland, 2023; Volume 37, ISBN 978-92-5-137966-0.
3. Lippe, R.S.; Cui, S.; Schweinle, J. Estimating Global Forest-Based Employment. *Forests* **2021**, *12*, 1219. [CrossRef]
4. Forestry Agency Annual Report on Forest and Forestry in Japan Fiscal Year. 2023. Available online: https://www.rinya.maff.go.jp/j/kikaku/hakusyo/r5hakusyo/attach/pdf/index-3.pdf (accessed on 28 August 2025).
5. Mozuna, M.; Yamaguchi, H.; Suzuki, H.; Yamaguchi, S.; Muneoka, H.; Sasaki, T.; Usui, K.; Iizawa, T.; Ohigashi, F.; Abe, K.; et al. Research on Unmanned Logging Operation by Automated Traveling Forwarder Using Electromagnetic Induction. *Bull. For. For. Prod. Res. Inst.* **2021**, *20*, 19–28. [CrossRef]
6. Nakazawa, M.; Otsuka, D.; Taki, S.; Yoshida, C.; Muneoka, H.; Iwaoka, M.; Nakashima, Y.; Matsui, A. Development of an Automatic Driving Forwarder Using in Combination with LiDAR-SLAM and GNSS. In Proceedings of the Proceedings of 55th International Symposium on Forestry Mechanization 7th Forest Engineering Conference, Firenze, Italy, 20 September 2023; p. 56. Available online: https://www.formec.org/images/proceedings/2023/Book_of_abstract_FORMEC_FEC_23.pdf (accessed on 24 January 2025).
7. La Hera, P.; Mendoza-Trejo, O.; Lindroos, O.; Lideskog, H.; Lindbäck, T.; Latif, S.; Li, S.; Karlberg, M. Exploring the Feasibility of Autonomous Forestry Operations: Results from the First Experimental Unmanned Machine. *J. Field Robot.* **2024**, *41*, 942–965. [CrossRef]
8. Lee, H.-S.; Kim, G.-H.; Ju, H.-S.; Mun, H.-S.; Oh, J.-H.; Shin, B.-S. Global Navigation Satellite System/Inertial Navigation System-Based Autonomous Driving Control System for Forestry Forwarders. *Forests* **2025**, *16*, 647. [CrossRef]
9. Viktor, P.; Kiss, G. Sensors in Self-Driving Vehicles: A Detailed Literature Review and New Trends. *Sensors* **2026**, *26*, 2153. [CrossRef]
10. Li, Y.; Ibanez-Guzman, J. Lidar for Autonomous Driving: The Principles, Challenges, and Trends for Automotive Lidar and Perception Systems. *IEEE Signal Process. Mag.* **2020**, *37*, 50–61. [CrossRef]
11. Santo, A.; Heredia, E.; Viegas, C.; Valiente, D.; Gil, A. Ground Segmentation for LiDAR Point Clouds in Structured and Unstructured Environments Using a Hybrid Neural–Geometric Approach. *Technologies* **2025**, *13*, 162. [CrossRef]
12. Usui, K. Data Augmentation Using Image-to-Image Translation for Detecting Forest Strip Roads Based on Deep Learning. *Int. J. For. Eng.* **2021**, *32*, 57–66. [CrossRef]
13. Lee, H.-S.; Kim, G.-H.; Ju, H.S.; Mun, H.-S.; Oh, J.-H.; Shin, B.-S. Developing Forest Road Recognition Technology Using Deep Learning-Based Image Processing. *Forests* **2024**, *15*, 1469. [CrossRef]
14. Zhou, K.; Liu, Z.; Qiao, Y.; Xiang, T.; Loy, C.C. Domain Generalization: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 4396–4415. [CrossRef]
15. Zhao, W.X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. A Survey of Large Language Models. *arXiv* **2025**. [CrossRef]
16. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models Are Few-Shot Learners. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 1877–1901.
17. Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Amin, M.; Hou, L.; Clark, K.; Pfohl, S.R.; Cole-Lewis, H.; et al. Toward Expert-Level Medical Question Answering with Large Language Models. *Nat. Med.* **2025**, *31*, 943–950. [CrossRef]
18. Uhm, M.; Kim, J.; Ahn, S.; Jeong, H.; Kim, H. Effectiveness of Retrieval Augmented Generation-Based Large Language Models for Generating Construction Safety Information. *Autom. Constr.* **2025**, *170*, 105926. [CrossRef]

19. Nawaz, U.; Muhammad, A.; Gani, H.; Naseer, M.; Khan, F.S.; Khan, S.; Anwer, R. AgriCLIP: Adapting CLIP for Agriculture and Livestock via Domain-Specialized Cross-Model Alignment. In *Proceedings of the Proceedings of the 31st International Conference on Computational Linguistics*; Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B.D., Schockaert, S., Eds.; Association for Computational Linguistics: Abu Dhabi, UAE, 2025; pp. 9630–9639. Available online: https://aclanthology.org/2025.coling-main.644/ (accessed on 23 April 2025).
20. Du, S.Q.; Tang, S.J.; Wang, W.X.; Li, X.M.; Guo, R.Z. Tree-Gpt: Modular Large Language Model Expert System for Forest Remote Sensing Image Understanding and Interactive Analysis. In Proceedings of the International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Cairo, Egypt, 2 September 2023; Volume XLVIII-1/W2-2023, pp. 1729–1736. Available online: https://isprs-archives.copernicus.org/articles/XLVIII-1-W2-2023/1729/2023/ (accessed on 9 August 2024).
21. Tan, J.; Zhang, H.; Yang, J.; Liu, Y.; Zheng, D.; Liu, X. ForestryBERT: A Pre-Trained Language Model with Continual Learning Adapted to Changing Forestry Text. *Knowl.-Based Syst.* **2025**, *320*, 113706. [CrossRef]
22. Wu, S.; Qiao, Y.; He, S.; Zhou, J.; Wang, Z.; Li, X.; Wang, F. FireCLIP: Enhancing Forest Fire Detection with Multimodal Prompt Tuning and Vision-Language Understanding. *Fire* **2025**, *8*, 237. [CrossRef]
23. Wang, J.; Shi, E.; Hu, H.; Ma, C.; Liu, Y.; Wang, X.; Yao, Y.; Liu, X.; Ge, B.; Zhang, S. Large Language Models for Robotics: Opportunities, Challenges, and Perspectives. *J. Autom. Intell.* **2025**, *4*, 52–64. [CrossRef]
24. Usui, K. Estimation of Potential Risk for Active Safety Systems Using Large Language Models. *J. Jpn. For. Eng. Soc.* **2026**, *41*, 5–14. [CrossRef]
25. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models from Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*; Meila, M., Zhang, T., Eds.; PMLR: Cambridge, MA, USA, 2021; Volume 139, pp. 8748–8763. Available online: https://proceedings.mlr.press/v139/radford21a.html (accessed on 22 August 2024).
26. Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. LAION-5B: An Open Large-Scale Dataset for Training next Generation Image-Text Models. *arXiv* **2022**. [CrossRef]
27. LAION Releasing Re-LAION 5B: Transparent Iteration on LAION-5B with Additional Safety Fixes | LAION. Available online: https://laion.ai/blog/relaion-5b (accessed on 20 May 2025).
28. Gadre, S.Y.; Ilharco, G.; Fang, A.; Hayase, J.; Smyrnis, G.; Nguyen, T.; Marten, R.; Wortsman, M.; Ghosh, D.; Zhang, J.; et al. DataComp: In Search of the next Generation of Multimodal Datasets. *arXiv* **2023**. [CrossRef]
29. Challa, N.R.; Mohan, B.C.; Babu, M.L.; Pavan Kumar, L.N. Enhancing Forest Fire Detection Using Zero-Shot Learning: A LLAVA and CLIP Vision-Language Model Approach. *Multimed. Tools Appl.* **2026**, *85*, 187. [CrossRef]
30. Guo, Y.; Conia, S.; Zhou, Z.; Li, M.; Potdar, S.; Xiao, H. Do Large Language Models Have an English Accent? Evaluating and Improving the Naturalness of Multilingual LLMs. *arXiv* **2024**. [CrossRef]
31. Gupta, V.; Chowdhury, S.P.; Zouhar, V.; Rooein, D.; Sachan, M. Multilingual Performance Biases of Large Language Models in Education. *arXiv* **2025**. [CrossRef]
32. Ilharco, G.; Wortsman, M.; Wightman, R.; Gordon, C.; Carlini, N.; Taori, R.; Dave, A.; Shankar, V.; Namkoong, H.; Miller, J.; et al. OpenCLIP. 2021. Available online: https://doi.org/10.5281/zenodo.5143773 (accessed on 20 May 2025).
33. Xu, H.; Xie, S.; Tan, X.E.; Huang, P.-Y.; Howes, R.; Sharma, V.; Li, S.-W.; Ghosh, G.; Zettlemoyer, L.; Feichtenhofer, C. Demystifying CLIP Data. *arXiv* **2025**. [CrossRef]
34. Zhang, B.; Zhang, P.; Dong, X.; Zang, Y.; Wang, J. Long-CLIP: Unlocking the Long-Text Capability of CLIP. *arXiv* **2024**. [CrossRef]
35. Zhai, X.; Mustafa, B.; Kolesnikov, A.; Beyer, L. Sigmoid Loss for Language Image Pre-Training. *arXiv* **2023**. [CrossRef]
36. Beyer, L.; Zhai, X.; Kolesnikov, A. Big Vision. Available online: https://github.com/google-research/big_vision (accessed on 6 February 2026).
37. Visheratin, A. NLLB-CLIP—Train Performant Multilingual Image Retrieval Model on a Budget. *arXiv* **2023**. [CrossRef]
38. Ministry of Health; Labour and Welfare. Shokuba No Anzen Site (Workplace Safety Website). Available online: https://anzeninfo.mhlw.go.jp/user/anzen/tok/anst00.html (accessed on 2 February 2026).
39. Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Kolesnikov, A.; et al. The Open Images Dataset V4: Unified Image Classification, Object Detection, and Visual Relationship Detection at Scale. *Int. J. Comput. Vis.* **2020**, *128*, 1956–1981. [CrossRef]
40. Zhang, X.; Li, S.; Hauer, B.; Shi, N.; Kondrak, G. Don't Trust ChatGPT When Your Question Is Not in English: A Study of Multilingual Abilities and Types of LLMs. *arXiv* **2023**. [CrossRef]
41. Nayak, S.; Jain, K.; Awal, R.; Reddy, S.; Steenkiste, S.V.; Hendricks, L.A.; Stanczak, K.; Agrawal, A. Benchmarking Vision Language Models for Cultural Understanding. In *Proceedings of the Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*; Al-Onaizan, Y., Bansal, M., Chen, Y.-N., Eds.; Association for Computational Linguistics: Miami, FL, USA, 2024; pp. 5769–5790. Available online: https://aclanthology.org/2024.emnlp-main.329/ (accessed on 2 June 2026).

42. Kim, E.; Shim, K.; Chang, S.; Yoon, S. Semantic Token Reweighting for Interpretable and Controllable Text Embeddings in CLIP. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024*; Al-Onaizan, Y., Bansal, M., Chen, Y.-N., Eds.; Association for Computational Linguistics: Miami, FL, USA, 2024; pp. 14330–14345. Available online: https://aclanthology.org/2024.findings-emnlp.837/ (accessed on 5 February 2026).
43. Zhou, K.; Yang, J.; Loy, C.C.; Liu, Z. Learning to Prompt for Vision-Language Models. *Int. J. Comput. Vis.* **2022**, *130*, 2337–2348. [CrossRef]
44. Ehrlich-Sommer, F.; Eberhard, B.; Holzinger, A. ForestGPT and Beyond: A Trustworthy Domain-Specific Large Language Model Paving the Way to Forestry 5.0. *Electronics* **2025**, *14*, 3583. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.