SAFIRE: Safety-Critical Benchmark for Fine-grained Fire and Smoke Understanding in Multimodal LLMs

Pengfei Li[1], **Naufal Suryanto**[1], **Sicheng Zhang**[1], **Mohammad Alsharid**[1], **Muzammal Naseer**[1,2]

[1]Department of Computer Science, Khalifa University, Abu Dhabi, UAE
[2]University of Western Australia, Australia
muhammadmuzammal.naseer@ku.ac.ae

Abstract

Multimodal Large Language Models (MLLMs) show strong progress on vision–language tasks, yet their reliability in safety-critical settings remains underexplored. Fire-smoke understanding is central to public safety and disaster response, but most existing benchmarks lack diverse real-world scenarios and context-aware evaluation. We introduce SAFIRE, a large-scale benchmark for fire-smoke understanding in MLLMs, comprising 83K captioned images from 20 scenarios and 193K multiple-choice VQA (MCVQA) generated from a 9.7K-image subset, spanning 10 evaluation dimensions from basic perception to higher-order reasoning. A GPT-5.4-assisted multi-stage verification pipeline with MLLM majority voting ensures annotation quality. Evaluating ten open-source MLLMs (8B–38B) yields an average accuracy of 61.9%, exposing major gaps in safety-critical reasoning. We further show that adapting vision encoders with only 7% of our domain-specific data boosts fire-scene classification accuracy from 20.1% to 64.5%, indicating that carefully curated data can yield substantial gains even when data volume is limited. All datasets, models, and code are available at https://risys-lab.github.io/SAFIRE/.

1 Introduction

Fire and smoke occur across diverse human and natural activities such as cooking, industry, celebrations, accidents, and wildfires. While their visual patterns often appear similar across settings (Chaturvedi et al., 2022), the underlying causes, risks, and appropriate responses vary widely. Understanding these phenomena requires not only visual recognition but also reasoning about context, intent, and consequence, which goes beyond conventional detection or segmentation. Nonetheless, most existing studies remain limited to narrow recognition tasks based on small, privately curated datasets with minimal contextual annotation (Han et al., 2024b; Zhu et al., 2025; Li et al., 2025a; Du et al., 2025; Boumaraf et al., 2025).

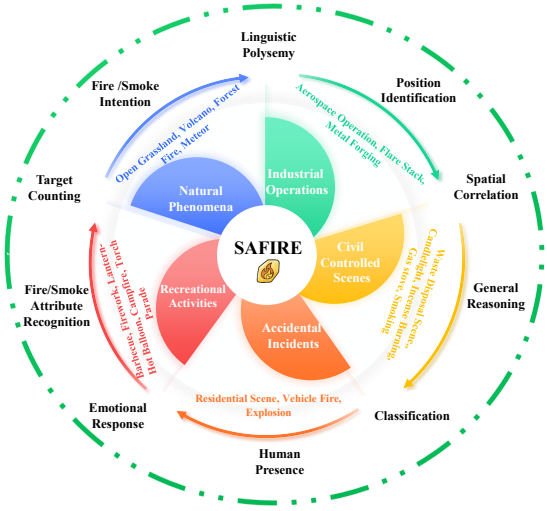

Figure 1: **From windmill center to the periphery:** SAFIRE consists of 5 processes spanning 20 real-world scenarios, with 10 evaluation dimensions and over 193K curated question-answer pairs derived from our subsets.

Recent MLLMs demonstrate strong zero-shot performance and richer contextual reasoning than traditional models (Liu et al., 2023; Bai et al., 2025b), yet existing fire-smoke benchmarks fall short in evaluating the situational reasoning required in safety-critical contexts (Table 1). They lack complex, context-aware queries such as determining whether visible smoke signals an emergency, leaving the reasoning limitations of current MLLMs insufficiently explored.

Figures 2 and 3 illustrate typical failure cases when contextual cues are ignored. In Fig. 2, water vapor is often misidentified as smoke during firefighting; in Fig. 3(a), warm lighting from bulbs or streetlights may trigger false fire alarms; and in Fig. 3(b), controlled pyrotechnic displays may appear as non-critical scenes. Broader scene cues such as firefighters, hoses, stages, or fire trucks help models resolve these ambiguities.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 33565–33589
October 24-29, 2026 ©2026 Association for Computational Linguistics

(a) Misleading "White Smoke" (b) Recognizing "No Smoke"

Figure 2: **Comparison of traditional and our dataset:** (a) False-positive "White Smoke" class image example from a local-perceptron dataset; (b) An image labeled "No Smoke" in our dataset, with a caption recognizing firefighters and a fire hose reel.

(a) Light Bulb on Wall (b) Stage Fire Performance

Figure 3: **Misclassifications without contextual reasoning:** (a) A warm-toned bulb triggers a false-positive "fire"; (b) A live fire show yields a false-negative "no fire." Incorporating global context, like identifying stage or fire trucks, corrects these mistakes.

These challenges underscore the need for context-aware, multimodal reasoning in fire and smoke understanding, yet progress remains limited by the absence of large-scale, high-quality datasets aligned with modern MLLM capabilities. To bridge this gap, we introduce **SAFIRE** (**S**afety-**A**ware **F**ire-smoke **I**nspection and **R**easoning **E**valuation), a large-scale benchmark for comprehensive, context-aware evaluation of MLLMs in fire-smoke reasoning. As shown in Fig. 1, SAFIRE comprises 83K captioned images from 20 real-world scenarios and 193K multiple-choice VQA pairs (from a 9.7K subset) spanning 10 evaluation dimensions, supporting both perceptual and higher-order reasoning at the scene level. Our main contributions are summarized as follows:

- We constructed *SAFIRE*, a multi-scenario benchmark explicitly designed to evaluate MLLMs' reasoning in fire-smoke contexts. Annotation quality is maintained through a GPT-5.4-assisted semantic verification pipeline combined with model-guided majority voting.
- *SAFIRE* enables systematic assessment of scene-level, contextual, causal, and behavioral reasoning, providing a testbed that extends beyond conventional classification or detection tasks. Evaluations on ten open-source MLLMs yield an average MCVQA accuracy of 61.9%, highlight remaining gaps in safety-critical reasoning.

2 Related work

Classical fire/smoke datasets. Early studies relied on handcrafted features and small binary datasets such as *BowFire* (Chino et al., 2015) and *F. Yuan* (Yuan et al., 2015), which focused on color and texture cues to distinguish fire or smoke from background scenes. With the advent of deep learning, larger labeled datasets (e.g., *Khan* (Khan et al., 2019), *L. He* (He et al., 2021), *EdgeFireSmoke* (Almeida et al., 2022)) enabled convolutional and attention-based models. Early datasets such as *FD* (Li et al., 2020) and *MAFireNet* (Khan et al., 2025) adopted binary fire/no-fire setups, later evolving toward finer-grained classification. Notably, *DFAN* (Yar et al., 2022) introduced scene-specific categories (e.g., vehicle- or building-related fires), though with limited per-class samples and semantic overlap, while *UFS-data* (Hosseini et al., 2022) proposed multi-label annotations for flame, white smoke, and black smoke. Despite these advances, video-derived frames and narrow scenario coverage restrict contextual diversity.

Detection, segmentation, and richer labels. Recent benchmarks have increasingly emphasized fire–smoke localization and multi-scenario annotations to enhance model robustness. Datasets such as *FiSmo* (Cazzolato et al., 2017), *Nemo* (Yazdi et al., 2022), *DFS* (Wu et al., 2023), and *HQFSD* (Qian et al., 2024) establish foundational localization pipelines, while newer benchmarks like *SmokeBench* (Qi et al., 2026) and *DetectiumFire* (Liu et al., 2026) enrich these with bounding-box annotations, fine-grained severity taxonomies, and curated "fire-like" negative samples to rigorously evaluate detection accuracy and visual reasoning. To mitigate the visibility degradation that often compromises fire smoke perception tasks, *SmokeBench*$_{(\mathrm{MM})}$ (Jin et al., 2025) contributes nearly 10K precisely aligned paired images specifically designed for early-stage fire desmoking, serving as a preprocessing foundation for downstream analysis. Furthermore, segmentation and sensor-augmented datasets have expanded the research landscape: *FireSentry* (Zhou et al.,

Table 1: **Comparison of SAFIRE with Previous Influential Works.** "Influential works" here refers to widely cited historical benchmarks (2015-2020) or foundational recent benchmarks (2021-2026). denotes Classification, denotes Detection, denotes Segmentation, denotes Sensor Data, denotes MCVQA and/or Captions. The proposed SAFIRE benchmark stands out for its overall scale, scenario diversity, and wide coverage.

Benchmark	Anno.	Fire	Smoke	# Classes	# Samples
BowFire (2015)		✓	✗	2	466
F. Yuan (2015)		✗	✓	2	21,329
Khan (2019)		✗	✓	4	72,012
FD (2020)		✓	✗	2	50,000
L. He (2021)		✗	✓	4	33,666
DFAN (2022)		✓	✗	12	3,084
EdgeFireSmoke (2022)		✓	✓	8	86,775
UFS-data (2022)		✓	✓	8	849,640
MS-FSDB (2024)		✓	✓	35	12,586
MAFire-Net (2025)		✓	✗	2	18,881
FiSmo (2017)		✓	✓	4	9,448
Nemo (2022)		✗	✓	4	6,702
DFS (2023)		✓	✓	3	9,462
HQFSD (2023)		✓	✓	4	12,166
S.Y. Kim (2023)		✗	✓	2	32,500
DetectiumFire (2026)		✓	✓	19	22,500
SmokeBench$_{(WACV)}$ (2026)		✗	✓	2	6,046
Mlich (2020)		✓	✗	1	6,386
WildfireSpread(2023)		✓	✗	2	13,607
MultiFire20K (2024)		✓	✓	4	20,500
UniInd-FireSmoke(2025)		✓	✓	6	30,101
FireSentry (2026)		✓	✓	2	20,710
SmokeBench$_{(MM)}$ (2026)	Paired Imgs	✗	✓	43	9,975
SAFIRE (Ours)		✓	✓	20	83K/9.7K/193K

Table 2: **SAFIRE dataset statistics across 20 fire-smoke relevant scenarios.** QA-I indicates the image subset used for generating questions (i.e. MCVQA-active images), and QA-P represents the total number of generated QA pairs.

Scenario	Full Set	QA-I (%)	QA-P
Accidental Incidents			
Residential Fire	4,280	469 (11.0%)	9,316
Explosion	3,924	463 (11.8%)	9,257
Vehicle Fire	1,616	175 (10.8%)	3,500
Industrial Operations			
Aerospace Operation	3,839	454 (11.8%)	9,056
Metal Forging	2,963	301 (10.2%)	6,006
Flare Stack	2,298	440 (19.1%)	8,799
Natural Phenomena			
Volcano	5,609	728 (13.0%)	14,561
Forest Fire	5,151	596 (11.6%)	11,822
Open Grassland	4,912	552 (11.2%)	11,013
Meteor	2,636	259 (09.8%)	5,115
Recreational Activities			
Barbecue	6,187	770 (12.4%)	15,400
Firework	6,105	749 (12.3%)	14,945
Campfire-Bonfire	5,560	702 (12.6%)	14,040
Torch Parade	2,922	300 (10.3%)	5,998
Sky Lantern-Hot Balloon	1,510	194 (12.8%)	3,880
Civil Controlled Scenes			
Waste Disposal	5,302	658 (12.4%)	13,159
Gas Stove	5,015	568 (11.3%)	11,321
Incense Burning	4,970	511 (10.3%)	10,212
Candlelight	4,043	431 (10.7%)	8,620
Smoking	4,000	348 (08.7%)	6,884
Total	**82,842**	**9,668 (11.7%)**	**192,904**

2026) delivers synchronized infrared–visible video streams and pixel-level fire masks for fine-grained wildfire spread forecasting, while resources like *WildfireSpread* (Gerard et al., 2023) and *UniInd-FireSmoke* (Li et al., 2025b) incorporate satellite or industrial sensor modalities.

Landscape summary. Table 1 summarizes 23 representative fire-smoke datasets (2015-2026). Classification tasks dominate (17/23), with fewer detection or segmentation resources. Only (2/23) include image-caption for evaluating reasoning, and none of the benchmarks include VQA pairs. Besides, over half (13/23) cover only one modality (fire or smoke), despite their frequent co-occurrence. Only a few—*MS-FSDB* (Han et al., 2024a), *DFAN* (Yar et al., 2022), *Detectium* (Liu et al., 2026), and *SmokeBench*$_{(MM)}$ (Jin et al., 2025)—offer relatively diverse scene coverage.

Rise of multimodal models. Recent advances in vision-language modeling, from encoder-only *CLIP* (Radford et al., 2021) to multimodal LLMs such as *GLM-4.1V* (Team et al., 2026b), *Qwen-3.5* (Qwen Team, 2026a), and *Gemma-3* (Team et al., 2025), enable zero-/few-shot transfer and contextual reasoning. However, prior benchmarks seldom touch the systematical evaluation of fire-smoke reasoning. **SAFIRE** addresses this gap by introducing diverse real-world scenarios, image-caption pairs, and MCVQA annotations that jointly assess perception, reasoning, and alignment within a unified multimodal framework.

3 SAFIRE Benchmark

We introduce **SAFIRE**, a large-scale benchmark for evaluating context-aware fire-smoke understanding in MLLMs. It contains 83K captioned-images across 20 real-world scenarios and a curated subset of 9.7K images with 193K multiple-choice VQA (MCVQA) pairs. Complementary image-caption annotations enable both perceptual understanding and higher-level reasoning. Figure 4 outlines the SAFIRE pipeline, which integrates data collection, dataset structuring, and multimodal evaluation into a unified testbed for fine-grained fire-safety reasoning and zero-shot assessment.

3.1 Data Collection and Structuring.

Images and captions are gathered from multiple sources and filtered through a multi-factor filtering (MFF) process combining automatic keyword matching, GPT-5.4-based semantic validation, and manual review (see Appendix B). The pipeline yields 83K high-quality multi-style samples, in-

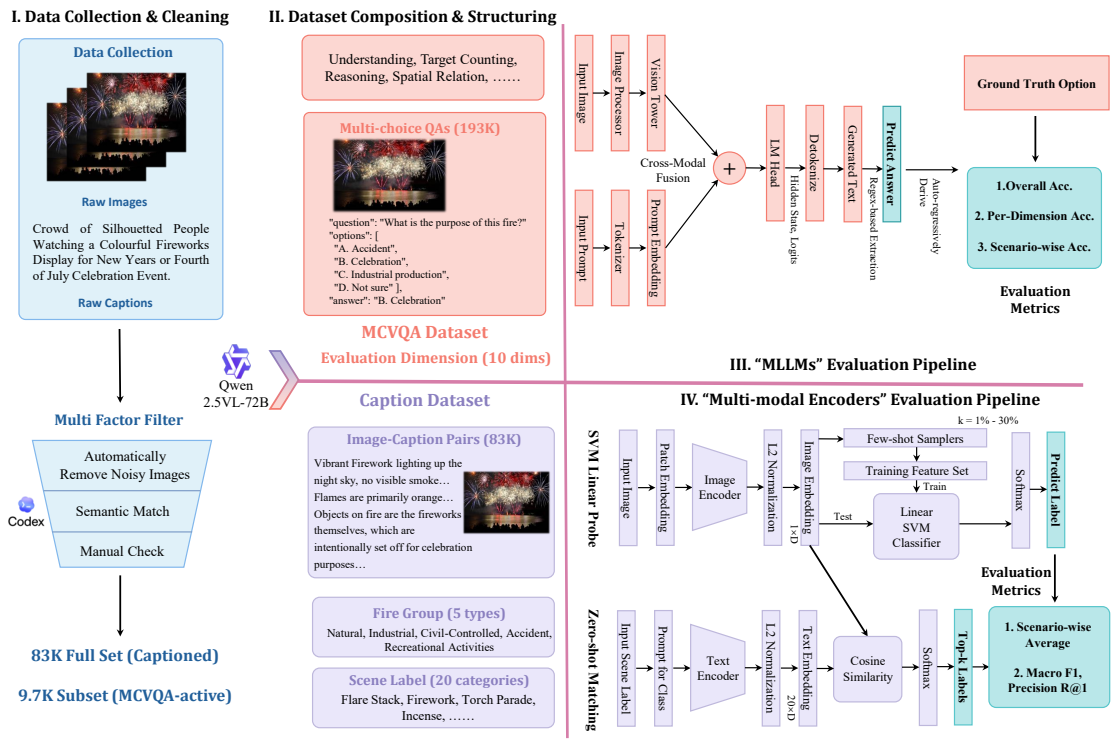

Figure 4: **SAFIRE pipeline.** Four stages: (I) multi-source collection and filtering; (II) construction of MCVQA and caption branches with hierarchical labels; (III) MLLM evaluation across ten reasoning dimensions; (IV) benchmarking of vision-language encoders under zero-shot and few-shot settings.

cluding both fire-present (positive) and minor fire-absent (negative) images across multiple scenarios, which helps suppress model overconfidence and habitual false-positive fire perception (Fig. 4, Stage I). Subsequently, 9.7K samples are selected to construct two subsets under a unified taxonomy: (i) an MCVQA set spanning ten reasoning dimensions, and (ii) an enhanced image set covering five groups and twenty scene labels (Fig. 4, Stage II).

Annotation Pipeline. All annotations are generated using a structured prompting framework built on Qwen2.5VL-72B-Instruct (Bai et al., 2025b). Standardized prompts define the expected outputs across the ten reasoning dimensions to ensure consistent question-answer generation. To reduce hallucinations, the model is instructed to "strictly follow the image content," and a "None of the above" option is included for ambiguous cases.

MCVQA Quality Control. Annotation quality is ensured through a multi-stage validation pipeline. First, we design scenario-specific prompts to explicitly inform the model of potential contextual biases (e.g., smoke dominance in Incense-burning scenes, or fire dominance in Firework and Metal Forging scenarios). All images, together with their generated MCVQAs, are then passed to GPT-5.4 for verification and correction. Subsequently, we employ majority voting across multiple MLLMs (Fig. 8) to identify inconsistent or ambiguous samples, which are further manually reviewed and corrected. This combination of automated validation and human verification provides reliable quality control. Finally, a human evaluation on 2,800 MCVQAs demonstrates an overall annotation accuracy of 87.1%, with scenario-wise accuracy ranging from 80% to 96%. More details about the quality control process, residual error types, and the effect of errors on model performance can be found in Appendix C.6 and D .

Visualization. Fig. 5 illustrates representative MCVQA samples, where each reasoning dimension is paired with images from distinct fire-smoke scenarios, highlighting SAFIRE's diversity and multimodal coverage. More details on dataset collection, cleaning, annotation, prompt design, and structuring are presented in Appendix B and C.

3.2 MLLM Evaluation.

We evaluate MLLMs as constrained generators (Fig. 4, Stage III). Each sample includes one image, its scenario label, and a multiple-choice question with four options (A-D). The prompt instructs the

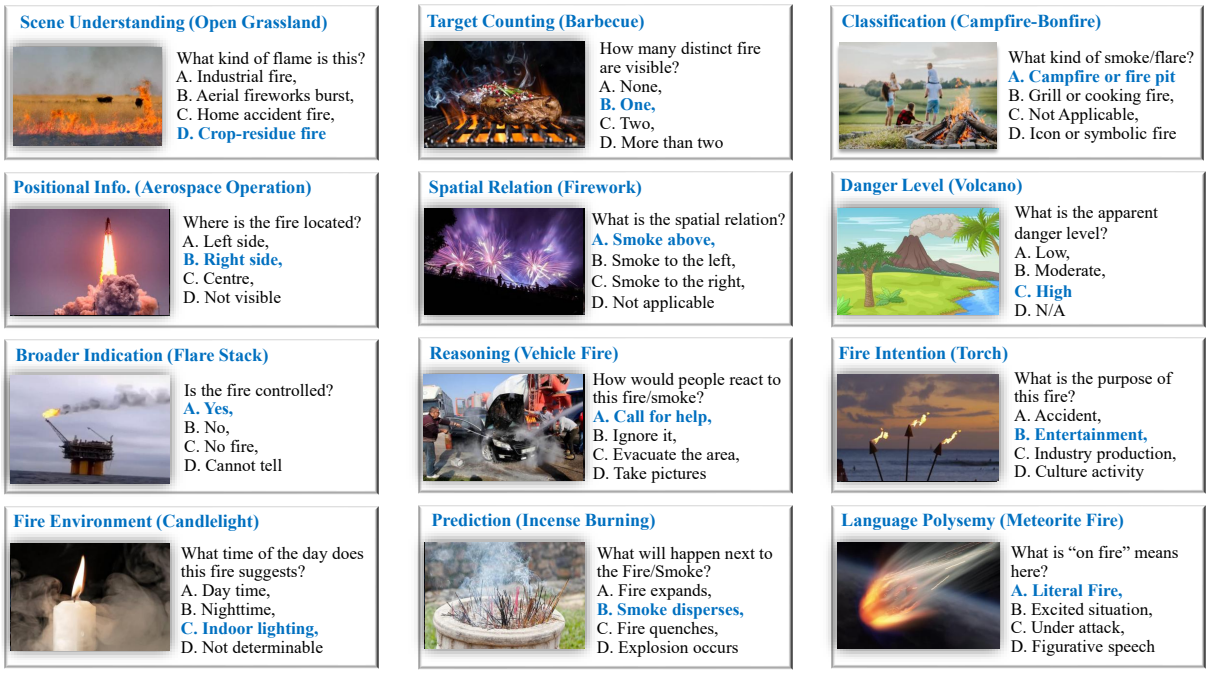

Figure 5: **Illustrative MCVQA samples.** Each question type (for example, scene understanding or target counting) is aligned with images from distinct fire-smoke scenarios.

model to respond with a single option letter, which is extracted and compared with the reference label. Questions test skills such as spatial awareness, object counting, intensity estimation, and semantic classification. Accuracy is reported along two axes: per-scenario (robustness across scenes) and per-dimension (skill-level performance). The detailed procedure is described in Appendix G.

3.3 Vision-Language Encoder Evaluation.

We evaluate underlying capabilities of vision-language encoders (e.g., CLIP (Radford et al., 2021)) for fire-smoke scene understanding using both zero-shot and few-shot protocols (Fig. 4 and Stage IV).

Zero-shot. Each class name (e.g., *campfire*) is inserted into an ensemble of prompt templates such as "A (*photo* or *image* or *scene*) of a campfire." We compute cosine similarity between frozen image and text features and select the label with the highest score.

SVM Linear Probe. A linear SVM is trained on frozen image embeddings using supervision ratios ranging from 1% to 30% with multiple random seeds for sampling, and evaluated on held-out samples. This setting examines how well the learned representations separate fire-related concepts under limited labeled data. Details of data splits, metrics, and training ratios are provided in Appendix G.

4 Experiments

4.1 Experimental Setup

We evaluate both MLLMs and vision-language encoders on the SAFIRE benchmark, as illustrated in Fig. 4. Stage III covers MLLM inference on the MCVQA task, while Stage IV evaluates vision-language encoders' performance. All MLLM experiments were conducted on an NVIDIA A100 multi-GPU cluster using the vLLM (Kwon et al., 2023) inference engine with automatic device mapping and `bfloat16` precision. For vision-language encoders, we used both pretrained general encoders and supervised CLIP+SVM baselines. Experiments are conducted on a curated 9.7K-image test subset. More details about model configurations can be found in Appendix H

4.2 MLLM Evaluation Results

We evaluate MLLMs in **Stage III** of the SAFIRE benchmark to assess their ability to solve multi-modal multiple-choice questions (MCVQA). This single-run evaluation measures model performance across ten reasoning dimensions and twenty real-world scenarios (under five processes) using structured QA samples. Metrics include overall accuracy, per-scenario performance, and per-dimension accuracy.

Performance Across Scenarios. Results on the twenty-scenario MCVQA benchmark (Table 3)

Table 3: **Accuracy (%) of MLLMs on the 20-scenario fire-smoke MCVQA benchmark. Bold** and underlined indicate the highest and second-highest accuracy in each **column**, and † denotes results from the thinking-enabled version of the model. Avg. is the overall average, Scene Avg. is the macro average.

Model	**Size**	**Natural Phenomena**				**Industrial Operations**			**Accidental Incidents**		
	(B)	Grassland	Volcano	Forest	Meteor	Aerospace	Flare Stack	Forging	Residential	Explos.	Vehicle
InternVL3.5	8	54.13	56.96	58.85	50.60	53.13	60.85	55.48	58.35	53.06	59.46
GLM4.1V	9	59.58	57.85	60.13	57.66	55.56	**75.37**	57.21	60.87	56.18	57.20
Qwen3.5	9	68.97	58.45	**63.97**	**68.28**	**62.37**	73.14	60.81	**65.17**	**60.73**	63.03
Qwen3.5†	9	64.30	57.50	61.43	59.09	59.46	74.50	56.06	61.52	56.58	60.66
Gemma-3	12	58.42	55.66	56.91	53.29	53.09	60.68	52.08	57.85	56.54	55.31
Gemma-3	27	58.24	55.49	60.23	53.67	52.11	66.17	54.35	57.98	54.25	58.57
Qwen3-VL	32	65.70	**61.67**	63.36	60.04	60.02	74.94	**62.29**	64.00	55.97	**65.66**
Qwen3-VL†	32	69.26	60.29	63.75	61.41	61.55	72.32	59.77	63.77	58.55	64.57
Qwen3.6	35A3	**69.32**	61.60	63.15	62.17	61.65	74.29	61.24	64.17	57.16	64.69
InternVL3.5	38	66.39	58.64	61.82	59.41	58.04	73.30	61.42	62.19	55.61	61.77
Scene Avg.	–	63.43	58.41	61.36	58.56	57.70	70.56	58.07	61.59	56.46	61.09

Model	**Size**	**Recreational Activities**					**Civil Controlled Scenes**					**All**
	(B)	Firework	SkyLantern	Barbecue	Campfire	Torch	Waste	Stove	Incense	Candle	Smoking	**Avg.**
InternVL3.5	8	54.39	62.14	55.75	65.75	70.27	60.91	67.87	46.82	65.86	60.34	58.37
GLM4.1V	9	52.09	57.85	50.26	60.08	66.85	59.63	70.51	58.13	69.39	60.68	59.70
Qwen3.5	9	62.18	70.38	57.55	69.93	75.02	61.03	**77.34**	57.29	70.96	64.13	64.89
Qwen3.5†	9	59.72	67.10	55.02	69.12	74.84	58.87	73.83	53.19	67.85	61.20	62.17
Gemma-3	12	56.44	56.52	48.56	59.11	69.42	58.29	62.30	50.18	56.80	50.60	56.31
Gemma-3	27	58.25	62.53	50.42	65.01	73.57	60.04	65.20	45.12	61.82	55.06	58.10
Qwen3-VL	32	**63.66**	70.54	61.10	70.02	73.87	63.97	73.64	61.31	**74.65**	**68.26**	65.35
Qwen3-VL†	32	63.29	**73.07**	**61.23**	69.48	76.44	**65.73**	75.07	**62.14**	73.35	66.28	**65.67**
Qwen3.6	35A3	63.64	71.56	60.98	**70.51**	**76.92**	62.71	75.92	58.77	71.23	67.50	65.53
InternVL3.5	38	54.55	68.02	57.99	68.85	73.96	63.71	70.67	54.48	73.28	59.15	62.65
Scene Avg.	–	58.82	65.97	55.89	66.79	73.12	61.49	71.23	54.74	68.52	61.32	–

show that most state-of-the-art vision-language models achieve modest performance, with an average accuracy of 61.87%. This indicates that considerable room remains for improvement in multimodal reasoning under safety-critical conditions. Increasing model size generally improves accuracy, but gains diminish beyond a certain scale; for example, InternVL 38B outperforms its 8B counterpart by around four points, suggesting that scaling alone is insufficient without domain adaptation. Performance also varies widely by scenario: natural phenomena settings (e.g., volcano, meteor) yield lower accuracies, whereas many controlled or familiar scenes (e.g., gas stove, candlelight) perform significantly better. This disparity likely stems from both data imbalance and limited exposure to rare phenomena imagery. Overall, SAFIRE reveals critical weaknesses in current MLLMs and offers guidance for developing models that generalize beyond everyday visual contexts.

Performance Across Reasoning Dimensions. Table 4 compares accuracies across ten evaluation dimensions. Performance varies substantially by task type. Structural dimensions such as *Classification* and *Human Presence* reach higher accuracies (76.4% and 90.3%), showing that MLLMs handle basic perceptual tasks well. In contrast, tasks like *Emotional Response* (61.7%), *Spatial Correlation* (42.9%), and *General Reasoning* (53.3%) remain challenging, highlighting persistent difficulty in understanding causality and semantics. Position Identification also lags behind, which is only 57.6% in average, indicating that models struggle to localize fires under occlusion or subtle conditions. Interestingly, Qwen3-VL† (32B) achieves higher accuracies than its succeeding and larger version (Qwen3.6 35B/A3B) across a few dimensions. Besides, tasks involving semantic ambiguity, such as *Polysemy*, show the widest gap (59.6%-71.6%), confirming that context disambiguation remains a major weakness. These results demonstrate SAFIRE's ability to reveal nuanced strengths and weaknesses that aggregate metrics obscure, emphasizing the value of dimension-wise evaluation for guiding targeted model improvement.

Effect of Reasoning-Enhanced Models. We further evaluate reasoning-enhanced ("thinking") models to enable a direct comparison with standard

Table 4: **Accuracy (%) across 10 evaluation dimensions on SAFIRE. Bold** indicates the highest accuracy per **dimension**, while underlined indicates the second-highest accuracy, and † denotes results from the thinking-enabled version of the model. Please go to Fig. 5 to find typical questions for each dimension.

Model	Target Counting	Class.	General Reasoning	Fire/Smoke Intention	Emotional Response	Polysemy	Fire/Smoke Attributes	Spatial Corr.	Position Identification	Human Presence
InternVL3.5 (8B)	67.40	73.12	49.84	67.75	54.73	59.56	58.95	34.18	55.63	88.20
GLM4.1V (9B)	70.12	74.52	48.87	68.32	54.24	60.60	63.02	29.38	60.95	92.85
Qwen3.5 (9B)	66.41	77.29	55.81	68.50	**70.69**	70.68	65.80	48.83	61.13	**93.07**
Qwen3.5† (9B)	67.78	76.38	52.32	66.85	66.34	68.13	63.76	45.57	52.14	91.06
Gemma-3 (12B)	42.21	71.75	49.76	71.43	45.24	66.39	56.13	38.08	54.98	85.09
Gemma-3 (27B)	48.87	76.05	50.66	69.82	49.02	63.06	57.73	42.48	56.58	87.71
Qwen3-VL (32B)	67.34	79.20	55.89	**73.76**	69.63	**71.61**	65.98	48.23	60.59	92.15
Qwen3-VL† (32B)	66.78	**81.85**	57.58	69.64	69.79	64.54	**66.54**	**53.77**	58.50	90.49
Qwen3.6 (35B/A3B)	**70.70**	76.30	**57.87**	70.82	69.90	65.04	66.38	49.59	**61.16**	91.62
InternVL3.5 (38B)	63.49	77.40	54.38	71.93	67.39	66.51	63.76	38.90	54.32	90.58
Avg.(%)	63.11	76.39	53.30	69.88	61.70	65.61	62.80	42.90	57.60	90.28

foundation models. As shown in Table 3, even the latest reasoning-oriented architectures achieve only around 62–66% overall accuracy, indicating that SAFIRE remains a challenging and discriminative benchmark. It effectively captures both performance gaps across models and the incremental gains brought by enhanced reasoning, particularly in fine-grained, safety-critical fire–smoke scenarios. Overall, explicit reasoning yields only modest gains and can even lead to performance degradation. The 9B thinking model is degraded from the corresponding foundation model by 2.7 percentage points, while the 32B series takes a 0.3-percentage-point gain. These results suggest that, despite advances in reasoning capabilities, current MLLMs continue to face significant challenges in achieving robust multimodal reasoning in safety-critical fire–smoke environments.

Evaluation Robustness and Implications.

Across scenarios and reasoning dimensions, performance gaps among models reflect the combined effects of model scale, architectural design, reasoning capability, and training data quality. Although the evaluated Qwen-family models belong to the same model family as the annotation generator and achieve relatively stronger performance, our expanded generator-family analysis shows that this effect is limited compared with the broader human–model gap; details are provided in Appendix I. In addition, we use the manually audited 2,800-MCVQA subset as a higher-confidence robustness check. The subset is not intended to isolate the causal effect of label correction, since it may differ from the full benchmark in composition and difficulty. Instead, it allows us to examine whether relative model trends remain broadly consistent under cleaner labels. Experimental results suggest that residual annotation noise does not change the main findings. We further include a scenario-prior ablation study in Appendix F, demonstrating that the evaluation results are mainly driven by image-grounded reasoning rather than by models exploiting scenario information as shortcuts. A supervised fine-tuning study is also provided in Appendix J, showing that SAFIRE can provide effective training supervision, with Qwen3-SAFIRE-SFT improving over the Qwen3-VL-8B baseline by 7.81 percentage points on a held-out image-level split. Benefiting from its rich annotations and hierarchical structure, SAFIRE enables fine-grained analysis and provides insights for future research directions, such as domain-specific pretraining, targeted data augmentation for rare or ambiguous scenarios, and tighter integration of visual and textual reasoning. Overall, SAFIRE establishes a comprehensive benchmark for developing reliable, context-aware multimodal systems for safety-critical fire and smoke understanding.

4.3 Multimodal Encoder Evaluation Results

Stage IV of the SAFIRE benchmark evaluates CLIP-based models under both zero-shot inference and supervised linear probing (CLIP+SVM). This single-run analysis examines how well vision-language encoders generalize to diverse fire-smoke scenarios and the effectiveness of lightweight domain adaptation over our high-quality data. Experiments span four zero-shot models and five super-

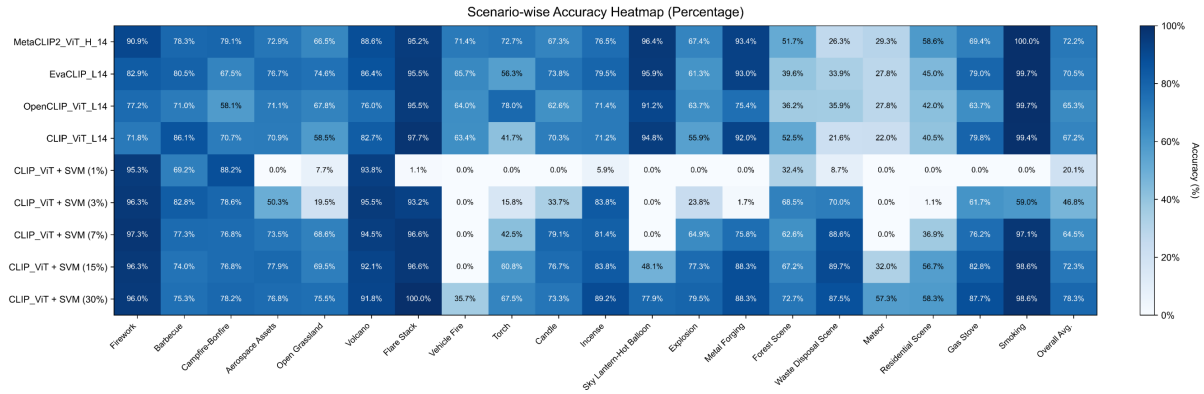

Figure 6: **Zero-shot accuracy of four multi-modal encoders** (i.e., CLIP-based models) across 20 scenarios in the SAFIRE benchmark. **The last five rows show results for CLIP+SVM**, which adds a simple tunable classification head. A random guess across 20 classes yields 5% accuracy.

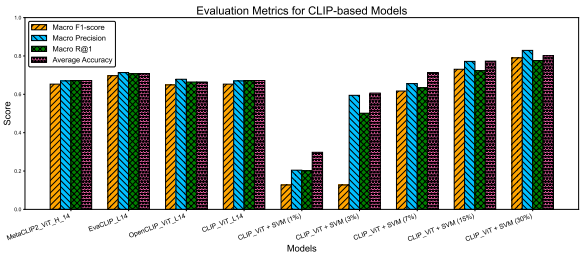

Figure 7: **Macro performance metrics (F1, Precision, Recall@1), and average accuracy across nine CLIP-based models.** The first four are zero-shot variants, while the latter are CLIP+SVM models with increasing label supervision.

vised variants across twenty scenarios (Fig. 6, 7).

Scenario-level Analysis. Fig. 6 visualizes scenario-wise accuracies. Zero-shot variants such as `MetaCLIP2` (Chuang et al., 2025) and `EvaCLIP` (Sun et al., 2023) substantially outperform the 5% random baseline and reach up to 72.2% accuracy. Highly salient categories such as *Smoking*, *Flare Stack*, and *Volcano* exceed 80% accuracy, while visually cluttered or ambiguous scenes such as *Meteor*, *Residential Scene*, and *Waste Disposal Scene* show weaker results. Model-wise variation is also evident: *Firework* achieves over 90% accuracy in `MetaCLIP` but only 71% in `CLIP_ViT_L14`, suggesting reliance on global scene cues.

Metric-level Analysis. Fig. 7 compares macro-level performance across nine encoders. Among zero-shot models, `EvaCLIP` achieves the best balance (F1=0.70, Recall@1=0.71, Precision=0.71), while other generic models lag behind. Although `MetaCLIP` attains the highest overall accuracy, its macro-level metrics remain consistently lower than those of `EvaCLIP`, indicating weaker class-balanced generalization.

Supervised variants (CLIP+SVM) show consistent gains as label supervision increases. With 30% supervision, CLIP+SVM achieves F1=0.79 and Precision=0.83, while even 7% supervision improves performance to F1=0.63—approaching most zero-shot baselines. These results highlight the efficiency of lightweight fine-tuning for specialized domains.

Summary and Implications. SAFIRE provides a challenging yet structured benchmark for vision-language alignment. Despite the difficulty of the 20-class classification task, strong zero-shot models such as `MetaCLIP2` reaches 72.2% accuracy, confirming the dataset's semantic separability and visual grounding. Overall, SAFIRE's diversity and granularity make it also a robust testbed for classification and retrieval-style evaluation under both zero- and linear probe settings, emphasizing the potential of lightweight adaptation for improving encoder performance in fire-smoke understanding.

5 Conclusion

We presented **SAFIRE**, a large-scale benchmark for context-aware fire-smoke understanding. The dataset comprises over **83K**-captioned images across **20** real-world scenarios and **193K** multiple-choice VQA pairs, organized under five groups (natural, industrial, accidental, recreational, civil controlled). SAFIRE supports two complementary evaluations: multi-dimensional reasoning with MLLMs via MCVQA, and zero-/few-shot scene classification with vision-language encoders. On SAFIRE, ten open-source MLLMs average **61.9%** MCVQA accuracy, while CLIP-style encoders

reach up to **72.2%** in zero-shot classification; adapting vision encoders with only 7% supervised data boosts fire-scene classification accuracy from 20.1% to 64.5%. These results highlight remaining gaps in safety-critical reasoning and the value of lightweight domain adaptation. We hope SAFIRE serves as a standardized testbed for building more reliable multimodal systems in fire- and smoke-related scenarios.

Limitations

While SAFIRE provides a strong foundation for fire and smoke understanding, several aspects offer opportunities for further improvement.

First, retrieval and alignment evaluations rely on limited natural-language prompts (e.g., "A (*photo* or *image* or *scene*) of a Vehicle fire"). Exploring more diverse prompts, including syntactic and semantic variations, as well as prompt tuning or renaming strategies (e.g., replacing "waste disposal fire" with "dumpster and garbage fire"), may further improve generalization and text–image alignment in safety-critical scenarios.

Second, the evaluated models are pretrained on general-domain corpora. Our adaptation is limited to the linear probe and one supervised fine-tuning configuration; broader evaluation across architectures and adaptation methods remains future work.

Finally, most annotations are generated by a vision-language model and refined through GPT-5.4 and small-model voting, with human verification for flagged samples. While this pipeline enables scalable and consistent data creation, biases may still persist due to the model's capability at the time of the Benchmark's design. Ongoing auditing (e.g., use different SOTA models to vote or correct question-answer pairs) and community-driven validation are needed to ensure fairness and reliability.

Overall, these limitations point to clear directions for future work, and addressing them will further strengthen SAFIRE as a reliable benchmark for safety-critical multimodal systems.

Acknowledgements

This research was funded by Khalifa University of Science and Technology through the Faculty Start-Ups under Project ID: KU-INT-FSU-2005-8474000775.

References

Jefferson Silva Almeida, Chenxi Huang, Fabrício Gonzalez Nogueira, Surbhi Bhatia, and Victor Hugo C de Albuquerque. 2022. Edgefiresmoke: A novel lightweight cnn model for real-time video fire–smoke detection. *IEEE Transactions on Industrial Informatics*, 18(11):7889–7898.

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, Yifei Shen, Yuanhan Zhang, Kaichen Zhang, Wenkang Zhang, Zheng Cheng, Nansen Zhang, Chunsheng Wu, Chunjiang Ge, Zimin Ran, and 11 others. 2026. Llava-onevision-2: Towards next-generation perceptual intelligence. *Preprint*, arXiv:2605.25979.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

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. 2025b. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Said Boumaraf, Muaz Al Radi, Fares Oussama Abdelhafez, Pengfei Li, Khalid Yousef Al Awadhi, Hamad Karki, Sahraoui Dhelim, and Naoufel Werghi. 2025. Vision-based air-flow monitoring in industrial flares system design using deep convolutional neural networks. *Expert Systems with Applications*, page 126733.

Mirela T Cazzolato, Letricia Avalhais, Daniel Chino, Jonathan S Ramos, Jessica A de Souza, Jose F Rodrigues-Jr, and A Traina. 2017. Fismo: A compilation of datasets from emergency situations for fire and smoke analysis. In *Brazilian symposium on databases-SBBD*, pages 213–223. SBC Uberlândia, Brazil.

Shubhangi Chaturvedi, Pritee Khanna, and Aparajita Ojha. 2022. A survey on vision-based outdoor smoke detection techniques for environmental safety. *ISPRS Journal of Photogrammetry and Remote Sensing*, 185:158–187.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829.

Daniel YT Chino, Letricia PS Avalhais, Jose F Rodrigues, and Agma JM Traina. 2015. Bowfire: detection of fire in still images by integrating pixel color

and texture analysis. In *2015 28th SIBGRAPI conference on graphics, patterns and images*, pages 95–102. IEEE.

Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, James R. Glass, LIFEI HUANG, Jason E Weston, Luke Zettlemoyer, Xinlei Chen, Zhuang Liu, Saining Xie, Wen tau Yih, Shang-Wen Li, and Hu Xu. 2025. Meta CLIP 2: A worldwide scaling recipe. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George Stoica, Winson Han, and 2 others. 2026. Molmo2: Open weights and data for vision-language models with video understanding and grounding. *Preprint*, arXiv:2601.10611.

Liwei Du, Zhihong Xu, Jianhong Gao, and Duanyu Chen. 2025. Firemultiformer: A transformer-based multimodal approach for predicting arc fault fires with fire images and fault signals. *IEEE Transactions on Instrumentation and Measurement*.

Sebastian Gerard, Yu Zhao, and Josephine Sullivan. 2023. Wildfirespreadts: A dataset of multi-modal time series for wildfire spread prediction. *Advances in Neural Information Processing Systems*, 36:74515–74529.

Xiaoyi Han, Nan Pu, Zunlei Feng, Yijun Bei, Qifei Zhang, Lechao Cheng, and Liang Xue. 2024a. Benchmarking multi-scene fire and smoke detection. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 203–218. Springer.

Xiaoyi Han, Yanfei Wu, Nan Pu, Zunlei Feng, Qifei Zhang, Yijun Bei, and Lechao Cheng. 2024b. Fire and smoke detection with burning intensity representation. In *Proceedings of the 6th ACM International Conference on Multimedia in Asia*, pages 1–8.

Lijun He, Xiaoli Gong, Sirou Zhang, Liejun Wang, and Fan Li. 2021. Efficient attention based deep fusion cnn for smoke detection in fog environment. *Neurocomputing*, 434:224–238.

Ali Hosseini, Mahdi Hashemzadeh, and Nacer Farajzadeh. 2022. Ufs-net: A unified flame and smoke detection method for early detection of fire in video surveillance applications using cnns. *Journal of Computational Science*, 61:101638.

Ailin Huang, Chengyuan Yao, Chunrui Han, Fanqi Wan, Hangyu Guo, Haoran Lv, Hongyu Zhou, Jia Wang, Jian Zhou, Jianjian Sun, Jingcheng Hu, Kangheng Lin, Liang Zhao, Mitt Huang, Song Yuan, Wenwen Qu, Xiangfeng Wang, Yanlin Lai, Yingxiu Zhao, and 74 others. 2026. Step3-vl-10b technical report. *Preprint*, arXiv:2601.09668.

Wenzhuo Jin, Qianfeng Yang, Xianhao Wu, Hongming Chen, Pengpeng Li, and Xiang Chen. 2025. Smokebench: A real-world dataset for surveillance image desmoking in early-stage fire scenes. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 12722–12728.

Salman Khan, Khan Muhammad, Shahid Mumtaz, Sung Wook Baik, and Victor Hugo C de Albuquerque. 2019. Energy-efficient deep cnn for smoke detection in foggy iot environment. *IEEE Internet of Things Journal*, 6(6):9237–9245.

Taimoor Khan, Zulfiqar Ahmad Khan, and Chang Choi. 2025. Enhancing real-time fire detection: An effective multi-attention network and a fire benchmark. *Neural Computing and Applications*, 37(18):11693–11707.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626.

Kang Li, Feiniu Yuan, and Chunmei Wang. 2025a. An effective multi-scale interactive fusion network with hybrid transformer and cnn for smoke image segmentation. *Pattern Recognition*, 159:111177.

Pengfei Li, Said Boumaraf, Muaz Al Radi, and Sicheng Zhang. 2025b. Uniind-firesmoke: Unified industrial fire-smoke dataset.

Songbin Li, Qiandong Yan, and Peng Liu. 2020. An efficient fire detection method based on multiscale feature extraction, implicit deep supervision and channel attention mechanism. *IEEE Transactions on Image Processing*, 29:8467–8475.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Zixuan Liu, Siavash H Khajavi, and Guangkai Jiang. 2026. Detectiumfire: A comprehensive multi-modal dataset bridging vision and language for fire understanding. *Advances in Neural Information Processing Systems*, 38.

Meta AI. 2024. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models.

Mistral AI. 2025. Mistral Small 3.1.

Tianye Qi, Weihao Li, and Nick Barnes. 2026. Smokebench: Evaluating multimodal large language models for wildfire smoke detection. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1043–1053. IEEE.

Jialong Qian, Chaoqun Hong, Kejie Zhang, and Jianglong Huang. 2024. High quality fire smoke dataset: A benchmark for fire and smoke detection. In *Proceedings of the 2nd International Workshop on Multimedia Content Generation and Evaluation: New Methods and Practice*, pages 26–35.

Qwen Team. 2026a. Qwen3.5: Towards native multimodal agents.

Qwen Team. 2026b. Qwen3.6-27B: Flagship-level coding in a 27B dense model.

Qwen Team. 2026c. Qwen3.6-35B-A3B: Agentic coding power, now open to all.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, and 304 others. 2026a. Gemma 4 technical report. *Preprint*, arXiv:2607.02770.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, and 74 others. 2026b. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *Preprint*, arXiv:2507.01006.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 56 others. 2025. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *Preprint*, arXiv:2508.18265.

Siyuan Wu, Xinrong Zhang, Ruqi Liu, and Binhai Li. 2023. A dataset for fire and smoke object detection. *Multimedia Tools and Applications*, 82(5):6707–6726.

Hikmat Yar, Tanveer Hussain, Mohit Agarwal, Zulfiqar Ahmad Khan, Suneet Kumar Gupta, and Sung Wook Baik. 2022. Optimized dual fire attention network and medium-scale fire classification benchmark. *IEEE Transactions on Image Processing*, 31:6331–6343.

Amirhessam Yazdi, Heyang Qin, Connor B Jordan, Lei Yang, and Feng Yan. 2022. Nemo: An open-source transformer-supercharged benchmark for fine-grained wildfire smoke detection. *Remote Sensing*, 14(16):3979.

Feiniu Yuan, Zhijun Fang, Shiqian Wu, Yong Yang, and Yuming Fang. 2015. Real-time image smoke detection using staircase searching-based dual threshold adaboost and dynamic analysis. *IET Image Processing*, 9(10):849–856.

Nan Zhou, Huandong Wang, Jiahao Li, Han Li, Yali Song, Qiuhua Wang, Yong Li, and Xinlei Chen. 2026. Firesentry: A multi-modal spatio-temporal benchmark dataset for fine-grained wildfire spread forecasting. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 2935–2946.

Wenyu Zhu, Shanwei Niu, Jixiang Yue, and Yangli Zhou. 2025. Multiscale wildfire and smoke detection in complex drone forest environments based on yolov8. *Scientific Reports*, 15(1):2399.

A Ethical Statement

All data used in this study were obtained either from publicly available sources or through experiments conducted by the authors. The SAFIRE benchmark was curated to advance research on safety-critical visual reasoning in fire and smoke scenarios and will be released for non-commercial research use. All pretrained models (e.g., Qwen-VL, InternVL) were employed in accordance with their respective open-source licenses. We used AI tool (e.g. Codex (GPT-5.4)) for data cleaning, and improving the language clarity of the main content; all conceptual and scientific content was solely verified by the authors.

B Data Collection and Cleaning

B.1 Dataset Collection

Images representing diverse fire and smoke scenarios were collected using a keyword-based search strategy across multiple web sources. To ensure high relevance and quality, we employed a multi-factor filtering pipeline throughout the collection process.

The first filtering step involved a strict keyword conjunction mechanism. For instance, to retrieve images for the "Open grassland fire" scenario, we used the combined query terms "grassland", "farm", "flatland" and "burning". Only images associated with all specified keywords were retained, thereby reducing semantic noise and increasing specificity. Additionally, a manually defined threshold parameter σ was introduced as a stopping criterion for automated webpage crawling. For example, in a web query structured as "*https://webname/search/term1&term2&term3/page=[1:σ]*", the crawler halted once the results began returning irrelevant content, preventing excessive inclusion of low-quality or unrelated samples.

Together, these two automated filters eliminated over 37% of irrelevant images prior to manual review, significantly improving the dataset's precision and reducing the overall cleaning burden in downstream stages.

B.2 Source Provenance

Table 5 summarizes the provenance of the raw candidate data before filtering. We combine published research datasets, public dataset hubs, public web platforms, authorized field data, and controlled combustion experiments to improve scenario diversity, with video-based sources being sampled into candidate frames before duplicate removal and semantic filtering. The controlled experiment set up was built in Sas Al Nakhl Campus of Khalifa University during Nov. 2023 to Mar. 2024 and the entire data collection protocol was approved by the university's Ethics Review Board. We reiterate that the resulting benchmark is for non-commercial research purposes only.

The raw provenance counts are approximate candidate records and may include similar images or low-relevance examples. After source-level filtering, semantic verification, and manual inspection, these candidates are reduced to the final 83K high-quality SAFIRE images described in Table 2.

B.3 Dataset Cleaning

Following data collection, we conducted a manual curation phase to refine the scenario selection. Each candidate scenario was assessed based on two main criteria: (1) overall image quality and noise level, and (2) semantic relevance to fire and smoke analysis. Several scenarios were discarded due to high false-positive rates or conceptual redundancy.

For example, the "Satellite Massive Fire" scenario was excluded due to frequent misclassifications involving laminar flow, haze, or cloud cover. Similarly, the following scenarios were removed for specific reasons:

- **Lightning Strike**: Predominantly returned synthetic computer-generated imagery (CGI) with low diversity and high semantic ambiguity (e.g., "lightning strike" may refer to military operations).
- **Welding Operation**: Often featured icons or close-up shots of machinery or workers, lacking sufficient visual focus on sparks or flames.
- **Electrical Fire**: Substantially overlapped with the *Home Fire* category, as most images depicted socket-related incidents already represented in the residential context.
- **Traditional Festival Rituals**: Visually resembled *Campfire* or *Bonfire* scenes, adding little scenario diversity.

After finalizing the 20 scenario categories and the initial data collection, we applied an **automation-manual combined cleaning process** across all retained images to remove remaining irrelevant or low-quality samples. This reduced the dataset from an initial 135.4K auto-collected images to 91K (after GPT-5.4 semantic match examination), and finally **83K high-quality images**

Table 5: **Source provenance of raw candidate data before filtering.** Counts are measured before the multi-factor filtering and manual inspection stages.

Source group	Source(s)	Raw data	Access
Published datasets	*BowFire* (Chino et al., 2015) (466); *UniInd-FireSmoke* (Li et al., 2025b) (3,789)	4,255 samples	Public research datasets. We cite the original sources and retain their research-use license constraints.
Roboflow datasets	15 public datasets, each containing 1,270–2,657 samples	25,879 samples	Public dataset sources.
Public platforms	YouTube (9,743); Shutterstock (9,356); Getty (9,589); TikTok (15,712); Bilibili (19,145); iStock (26,869)	90,414 images	Images or frames collected and filtered for non-commercial research use. SAFIRE is released for research purposes only.
Field data	Gas-flare samples	2,271 samples	Collected with authorized field access.
Controlled experiments	80 combustion experiments	About 12,600 frames	Collected by varying O_2, CH_4, C_2H_6 composition and flow velocity to generate diverse fire and smoke behaviors;

(after manual inspection).

As a result, only **61.2%** of the initial data was retained in the final SAFIRE benchmark. This rigorous filtering yields a cleaner, more diverse, and semantically coherent dataset, providing a solid foundation for robust vision-language research in fire and smoke understanding. SAFIRE is released for research use only, and a summary of final statistics and dataset categorization is provided in Table 2.

C Dataset Composition & Structuring

C.1 Annotation Structure and Benchmark Objectives

SAFIRE comprises two complementary annotation branches, each tailored to evaluate distinct multimodal reasoning capabilities:

- **Multi-Choice Visual Question Answering (MCVQA):** This branch is designed to assess the fine-grained reasoning abilities of MLLMs. It includes 193K questions spanning 10 reasoning dimensions (e.g., context, causality, behavior), targeting high-level understanding of fire and smoke scenes beyond simple recognition.
- **Image-Caption (IC) Pairs:** This branch supports the evaluation of multimodal encoders, particularly in zero-shot and retrieval settings. Each of the 83K images is paired with a carefully constructed caption capturing key semantic attributes, including actors, objects, fire-smoke states, and safety cues.

These branches enable comprehensive benchmarking across reasoning depth and modality alignment, supporting both autoregressive MLLMs and encoder-based architectures. Each image is also annotated with a fire category and scene label to facilitate fine-grained scenario analysis.

C.2 Prompting Strategy and Instruction Design

All annotations, including both MCVQA and IC pairs, are generated simultaneously using a structured prompting pipeline built on the Qwen2.5-VL-72B-Instruct (Bai et al., 2025b) vision language model. A standardized instruction defines the expected output format and covers 10 reasoning dimensions such as scenario type, spatial localization, fire-smoke interaction, and human activity.

To minimize hallucinations, the model is explicitly instructed to *"strictly follow the image content when generating questions and answers"*. In ambiguous situations, a "None of the above" option is included, allowing the model to abstain when none of the listed answers is correct. The following boxes illustrate the default system instruction and captioning guidelines used during annotation:

Default Instruction

"You are an advanced VLM tasked with producing detailed captions for fire-related images, each from one of 20 predefined fire-related scenarios..."

.....

"... Generate captions that thoroughly describe visible fire and smoke elements. Include sufficient details for the 10 evaluation dimensions...

Form ground-truth QA pairs to assess reasoning capabilities."

Captioning Guidelines

1. Context Awareness: Confirm the fire-related context and avoid symbolic or irrelevant images.
2. Visual Description: Describe key attributes of fire and smoke, surrounding context, and activity.
3. Output Format: Produce a structured JSON object with captions, key indicators, and MCVQA entries.

C.3 MCVQA Design and Reasoning Dimensions

Each selected MCVQA-active image is paired with approximately 20 multiple-choice questions targeting multi-level visual and contextual reasoning. These are categorized as follows:

QA Instructions (Q1-Q20)

Scene-Level Evaluation (Q1): Scenario classification, e.g., "What kind of fire/smoke is this?"
Spatial & Instance-Level (Q2-Q4): Fire position (left/center/right), number of fire sources, danger level.
Recognition (Q5-Q10): Smoke presence, volume, color, flame visibility, time of day, negative cases.
Reasoning & Context (Q11-Q20): Fire source, human presence, firefighting activity, trend prediction, emergency response, and semantic ambiguity.

C.4 Structured Output Format

All annotations conform to a unified JSON schema, consisting of the image ID, caption, key indicators, and MCVQA entries. This format ensures compatibility with downstream tools and promotes reproducibility in multimodal benchmarks.

Output JSON Skeleton

```
{
  "Image_name": "...", "caption": "...",
  "Key Indicators": {
    "pos": "...", "danger level": "...",
    "Fire trend": "..."
  },
  "QA_pairs": [
    {"Question 1": "...", "Options": [...],
    "Answer": "..."},
    {"Question 2": "...", "Options": [...],
    "Answer": "..."},
    ...
  ]
}
```

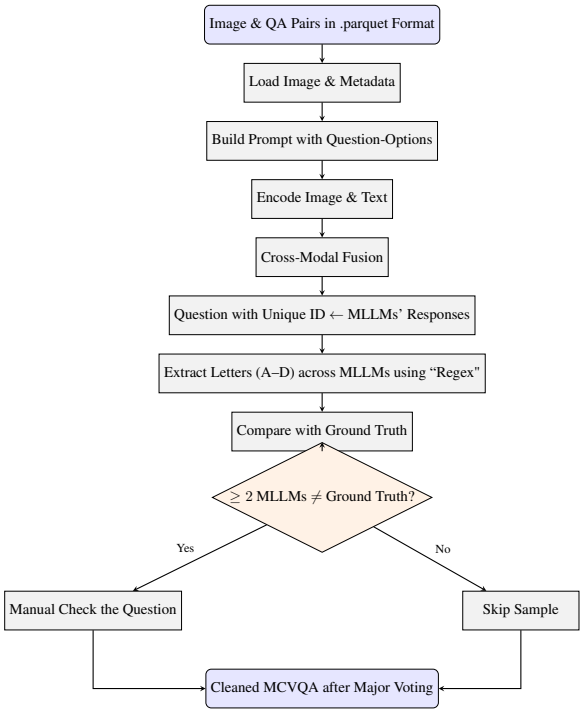

Figure 8: **Majority voting pipeline for cleaning multi-scenario MCVQA answers from multiple MLLMs.**

C.5 Handling MCVQA Failures

During caption generation, a subset of samples failed to yield valid JSON. These failures were categorized as: (1) **complete failures**, with no output returned, and (2) **partial failures**, where captions were generated but no MCVQA pairs. In partial cases, captions were retained and included in the CLIP branch, enriching the pool of aligned image-text pairs.

C.6 GPT-5.4 Verification and Majority Voting for MCVQA Correction

To ensure the quality of MCVQAs, we adopt a GPT-5.4-assisted cleaning pipeline combined with majority voting across multiple MLLMs.

The verification process consists of two rounds of refinement. In the first round, we design scenario-specific prompts to make the model aware of potential contextual biases and critical visual cues for each question. For example, in the *Firework* scenario, the model is explicitly instructed to recognize that fire may dominate the scene, care-

Table 6: **Residual annotation error analysis.** Error categories are computed over the 12.9% incorrect cases identified in the 2,800-sample human audit.

Error type	Percent (%)	Example
Fire/smoke visibility and visual attributes	24.5	`campf_654.jpg`: a pile of unlit wood was labeled as visible flame.
Semantic mismatch	20.8	`volca_308.jpg`: a white steam plume was labeled as molten lava glow.
Answer format error	17.2	`rocke_424.jpg`: the ground-truth text was "Emergency response", but it used the wrong option label.
Context inference error	14.2	`fores_557.jpg`: a firefighter wearing a black uniform in a burned forest was labeled as no people and no active fire, which is only partially correct.
Localization/counting error	9.9	`firew_213.jpg`: a cartoon of fireworks was labeled as having more than two real fire sources.
Real vs. symbolic error	8.4	`home_419.jpg`: a fire icon was labeled as literal fire.
Spatial relation error	5.1	`rocke_93.jpg`: an aircraft smoke trail was labeled as smoke rising above fire, while it actually trails behind the aircraft.

Table 7: **Inter-model consistency in majority voting.**

Votes	# Samples	Rate
3/3 same answer	101,332	52.53%
2/3 same answer	85,958	44.56%
3 different answers	5,614	2.91%

Table 8: **Model-wise agreement with the original answer.** No single verifier dominates the voting process.

Verifier model	Agreement
Mistral-Small-3.1-24B-Instruct	56.67%
Gemma-3-12B-it	55.46%
Llama-3.2-11B-Vision-Instruct	52.37%

fully count distinct bursts, and avoid over-counting spark fragments.

In the second round, GPT-5.4 performs logical consistency checking to identify contradictions between image content and generated answers. For instance, if an image contains only fire without visible smoke, questions regarding the "spatial relationship between fire and smoke" should be answered as *N/A*, and if the visible evidence indicates a fire with only fireplace/wood/bricks, then a "high danger" label is usually inconsistent and should be corrected.

After GPT-5.4 refinement, we further apply majority voting as a disagreement detector. As illustrated in Fig. 8, instead of using majority voting as an automatic relabeling mechanism, we use this method to signal potential incorrect data where most models fail. We use three non-Qwen verifier models that are independent of the Qwen annotation generator: LLaMA-3.2-11B, Mistral-3.1-24B, and Gemma-3-12B. For each question, model predictions (options A–D) are extracted using regular expressions and compared with the original annotated answer. Samples are flagged for manual review when the three verifiers produce no clear majority, or when the majority vote contradicts the annotated answer; otherwise, they are automatically accepted. This strategy efficiently prioritizes human inspection toward potentially flawed annotations while avoiding direct dependence on a single verifier.

As shown in Table 7, the three verifier models produce a clear majority in 97.09% of MCVQA samples. Only 2.91% of samples produce three different answers, indicating that complete inter-model disagreement is rare. The final manually double-reviewed set from this process consists of three mutually exclusive subsets: (1) cases where the three verifier models give three different answers (5,614 samples), (2) cases where the original annotated answer differs from the 2/3 majority vote (1,308 samples), and (3) cases where the original annotated answer differs from the 3/3 unanimous vote (215 samples). In total, 7,137 MCVQA samples, corresponding to 3.70% of the full MCVQA set, were double-reviewed through the majority-voting process.

Table 8 further reports model-wise agreement with the original annotated answers. The three agreement rates suggest that the voting procedure does not over-rely on a single dominant verifier model. Moreover, the pipeline can identify ambiguous questions where model responses fall outside the predefined options, indicating unclear or inconsistent labels that require further investigation.

Table 9: **Full-benchmark and manually audited subset accuracy.** Results compare model accuracy on the complete SAFIRE MCVQA set and the manually corrected 2,800-MCVQA set.

Model	Original Full-Set Acc.	Audited-Subset Acc.	Difference
Qwen3.5-9B non-thinking	64.89%	68.98%	+4.09
Step3-VL-10B	63.39%	67.96%	+4.57
Qwen3.5-9B thinking	62.17%	67.41%	+5.24
InternVL3.5-8B	58.37%	63.73%	+5.36
LLaVA-OneVision-2-8B	61.61%	64.30%	+2.69
GLM-4.1V-9B-Base	59.70%	62.90%	+3.20
Gemma-4-E4B-it	59.30%	60.60%	+1.30
Molmo2-8B	56.92%	60.29%	+3.37

D Annotation Error Analysis

Building on the human verification described in Appendix C.6, we further analyze the incorrect portion of the 2,800 manually audited MCVQA samples. Table 6 summarizes the main residual error types from the remaining 12.9% annotation error rate. Most errors are correctable visual-grounding ambiguities, semantic mismatches, or answer-format inconsistencies.

Robustness under manually audited labels. To assess whether residual annotation noise could affect the benchmark conclusions, we further evaluated an extended family of models on the manually audited 2,800-MCVQA subset. This subset was manually corrected after the annotation audit and provides a higher-confidence reference set for examining whether the observed model trends remain stable under cleaner labels. We use this analysis as a robustness check rather than a precise estimation of the isolated effect of label correction because the audited subset is not identical to the full benchmark. As shown in Table 9, models obtain accuracies that are 1.30 to 5.36 percentage points higher on the audited subset than on the full SAFIRE MCVQA benchmark. More importantly, the relative ordering of models remains broadly consistent, and all models still perform well below human reference performance. These results suggest that residual annotation noise may affect absolute scores, but it does not overturn the main conclusion that current MLLMs still struggle with SAFIRE's fire-smoke reasoning tasks.

E Dimension Correlation Analysis

To further analyze the structure of SAFIRE, we measure the association between different evaluation dimensions. We first compute correlations at the question-template level. Each answer is converted to its multiple-choice option label (A–D), and Cramér's V is computed for every pair of question templates. We then aggregate these scores to the dimension level: for any two dimensions, their correlation is the average Cramér's V across all question-template pairs drawn from the two dimensions. Values near 0 indicate weak association, while values near 1 indicate strong overlap.

Table 10: **Strongest and weakest cross-dimension associations.** Abbreviations: Class. = Classification, Emot. Resp. = People Emotional Response, Pos. Ident. = Position Identification, Intention = Fire/Smoke Intention, Human Pres. = Human Presence, Spatial Corr. = Fire–smoke Spatial Correlation.

Type	Dim. A	Dim. B	Mean V
High-1	Class.	Emot. Resp.	0.410
High-2	Counting	Pos. Ident.	0.407
High-3	Intention	Emot. Resp.	0.365
Low-1	Human Pres.	Pos. Ident.	0.030
Low-2	Human Pres.	Spatial Corr.	0.040
Low-3	Counting	Human Pres.	0.060

Table 10 selectively reports the strongest and weakest cross-dimension associations, where we can find even the strongest association is moderate: *Classification* and *People Emotional Response* reach 0.410, which is expected because the type of fire or smoke scene often influences how people are likely to react. The second strongest association, between *Target Counting* and *Position Identification* (0.407), is also intuitive because both rely on locating salient fire or smoke sources in the image. Similarly, *Fire/Smoke Intention* and *People Emotional Response* are associated (0.365), since controlled, accidental, and natural fire contexts imply different human reactions.

The weakest correlations are substantially lower. *Human Presence* has very weak association with

Table 11: Ablation study of scenario priors on the full 193K MCVQA set, where * denotes the original setting used in the main evaluation.

Input mode	Acc.
Question only	29.26%
Question + Scenario, without image	39.17%
Question + Image, without Scenario	**64.13%**
Question + Image + Scenario*	**64.89%**

Position Identification (0.030), *Fire–smoke Spatial Correlation* (0.040), and *Target Counting* (0.060), indicating that whether people appear in an image is largely independent of where fire or smoke is located and how many sources are visible. The full correlation analysis also shows low-to-moderate associations for other semantically distinct pairs, such as *Fire–smoke Spatial Correlation* and *General Reasoning* (0.190). Overall, these results indicate that SAFIRE dimensions are semantically related where expected, but remain sufficiently distinct to support multi-dimensional evaluation.

F Scenario-Prior Ablation Study

In the MCVQA evaluation, certain scenarios may be correlated with particular answers. For example, stove scenes may be associated with controlled cooking, whereas residential-fire scenes may be associated with danger or emergencies. Such scenario–answer correlations could provide models with non-evident priors, potentially allowing them to improve performance through prior-based shortcuts rather than image-grounded reasoning.

To probe this possibility, we conduct full-set ablations using Qwen3.5-9B model. We selectively remove the image and/or the scenario name while keeping the question format unchanged. This design separates the contributions of scenario information and visual evidence. As shown in Table 11, explicitly providing the scenario name yields a measurable prior when images are absent: the accuracy increases from 29.26% for question-only input to 39.17% for question-plus-scenario input, a gain of 9.91 percentage points. However, visual evidence is substantially more informative: adding the image to the question-only setting improves accuracy by 34.87 percentage points, from 29.26% to 64.13%. Moreover, removing scenario information from the original evaluation setting results in only a 0.76-percentage-point drop (64.89% to 64.13%). Therefore, although scenario information provides a measurable prior, the performance in our main evaluation protocol is primarily driven by image-grounded reasoning rather than by exploiting scenario information as a shortcut.

G Evaluation Pipeline Algorithms

The full procedural steps for the evaluation of vision-language models on the MCVQA Fire dataset are detailed in Algorithm 1, which outlines how each image is loaded from the provided root directory, pairing with respective scenario labels and question-specific prompts. Besides, it also illustrates how model predictions are generated under a maximum token limit T_{max}. The pipeline then extracts answer choices from model outputs to calculate corresponding scores.

For MCVQA scoring, answer extraction is performed on the model's final-answer field rather than by requiring the entire response to be a single letter. Specifically, our prompt asks models to provide the final choice in a machine-readable format, such as `<output>A</output>`. For thinking models, reasoning may appear before the final answer, but only the final-answer field is parsed. If the exact field is missing, we additionally recover unambiguous final-choice patterns such as '`Answer: A`', '`Option A`', or '`the option is A`'. Ambiguous or missing final answers are treated as invalid. This post-processing design ensures that the reported accuracy is not a simple, strict single-letter matching.

In Algorithm 2, the full procedural steps for CLIP-based fire scene classification under both zero-shot and linear probe settings are detailed. This algorithm outlines how class-specific textual prompts are tokenized and encoded into normalized text features using the pretrained CLIP text encoder, and indicates the way that image features are extracted and normalized for all samples in the dataset using the CLIP image encoder. Moreover, the algorithm also illustrates how the evaluation across multiple data regimes. Specifically, our algorithm computes zero-shot predictions via cosine similarity between image and text features, and additionally trains a linear SVM classifier on varying fractions (from 1% to 30%) of the labeled image features to test linear probe capabilities.

H Model Configuration

MLLM Configuration. We evaluated the following models: *Qwen3.5 (9B, with and without reason-*

Algorithm 1: Evaluation on MCVQA

Input: Vision-language model $\mathcal{M}$, dataset $\mathcal{D}$, image root $\mathcal{I}$, max tokens $T_{\max}$
Output: Accuracy Acc, per-dimension $\mathrm{Acc}_{\mathrm{dim}}$, per-scenario $\mathrm{Acc}_{\mathrm{scen}}$

Initialize accuracy counters $C, N \leftarrow 0$;
foreach $(s, x) \in \mathcal{D}$ **do**
  Load image $I \leftarrow \mathcal{I} + s$;
  if *I unreadable* **then**
    continue;
    end if;
  $sc \leftarrow x[\texttt{scenario}]$
  foreach $qa \in x[\texttt{QA_pairs}]$ **do**
    Prompt $p \leftarrow \texttt{build_prompt}(qa, sc)$;
    Input $z \leftarrow \texttt{processor}(I, p)$;
    if *image tokens missing* **then**
      continue;
      end if;
    Output $\hat{y} \leftarrow \mathcal{M}(z)$ with $T_{\max}$;
    Prediction $\hat{a} \leftarrow$ extract letter from $\hat{y}$;
    a $\leftarrow$ first letter of $qa[\texttt{answer}]$;
    $N \leftarrow N + 1$;
    if $\hat{a} = a$ **then**
      $C \leftarrow C + 1$;
      increment correct in $\mathrm{Acc}_{\mathrm{dim}}$, $\mathrm{Acc}_{\mathrm{scen}}$;
      end if;
    Increment total in $\mathrm{Acc}_{\mathrm{dim}}$, $\mathrm{Acc}_{\mathrm{scen}}$;
  end for;
end for;
$\mathrm{Acc} \leftarrow \frac{C}{N}$;
return Acc, $\mathrm{Acc}_{\mathrm{dim}}$, $\mathrm{Acc}_{\mathrm{scen}}$

Algorithm 2: CLIP-Based Fire Scene Classification

Input: Pretrained CLIP model $\mathcal{M}$, dataset $\mathcal{D}$, class names $\mathcal{C}$
Output: Zero-shot accuracy Acc_{zs}, few-shot accuracy Acc_{svm}, evaluation metrics

Tokenize prompts $\mathcal{T}_i \leftarrow$ "A photo of a c_i" for each $c_i \in \mathcal{C}$;
Compute normalized text features $F_t \leftarrow \mathcal{M}_{\text{text}}(\mathcal{T})$;
foreach *class c in $\mathcal{D}$* **do**
  Load images I_c;
  foreach *image $i \in I_c$* **do**
    Encode and normalize $f_i \leftarrow \mathcal{M}_{\text{image}}(i)$;
    Store f_i and label for both evaluation modes;
  end for;
end for;
/* Train classification head under different supervision ratios */
foreach *ratio* $r \in \{1\%, 3\%, 7\%, 15\%, 30\%\}$ **do**
  Split each class's features: $r\%$ for training, remainder for testing;
  Train SVM $\mathcal{S}_r$ on training set;
  Predict labels on test set using $\mathcal{S}_r$;
  Compute $\mathrm{Acc}_{svm}^{(r)}$, precision, recall, F1;
  Save results for comparison;
end for;
Predict zero-shot labels via $\arg\max \mathrm{softmax}(f_i \cdot F_t^{\top})$;
Compute Acc_{zs} and associated metrics;
Export all results to Excel;
return Acc_{zs}, $\mathrm{Acc}_{svm}^{(r)}$, *full metric logs*

ing) (Qwen Team, 2026a), *GLM-4.1V (9B)* (Team et al., 2026b), *Gemma-3 (12B, 27B)* (Team et al., 2025), *Qwen3-VL (32B, with and without reasoning)* (Bai et al., 2025a), *Qwen3.6 (35B-A3B)* (Qwen Team, 2026c), and *InternVL3.5 (8B, 38B)* (Wang et al., 2025). Besides, we used *Mistral-Small-3.1-Instruct (24B)* (Mistral AI, 2025), *Gemma-3-it (12B)* (Team et al., 2025), and *Llama-3.2-Vision-Instruct (11B)* (Meta AI, 2024) for majority voting. In the supplementary experiments, we further evaluated *Step3-VL (10B)* (Huang et al., 2026), *LLaVA-OneVision-2 (8B)* (An et al., 2026), *Gemma-4-E4B-it* (Team et al., 2026a), *Molmo2 (8B)* (Clark et al., 2026), *Qwen3.6 (27B)* (Qwen Team, 2026b).

All models were obtained from their official Hugging Face repositories under their respective licenses, and evaluated with default inference settings to ensure consistency and comparability across architectures.

Encoder Config. For Stage IV, we evaluate vision-language encoders using both pretrained general encoders and supervised CLIP+SVM baselines. Experiments are conducted on a curated

9.7K-image test subset. The representative pre-trained baselines include *CLIP* (Radford et al., 2021), *MetaCLIP2* (Chuang et al., 2025), *EVA-CLIP* (Sun et al., 2023) and *OpenCLIP* (Cherti et al., 2023).

I Bias Analysis on Generator-family Model and Human Performance

Since the SAFIRE MCVQA annotations were initially generated by Qwen2.5-VL-72B-Instruct, Qwen-family models may be favored by synthetic patterns inherited from the generator. To examine this concern, we additionally expanded the evaluation to include both Qwen-family and non-Qwen-family models under the same prompt, decoding, answer-extraction, and scoring protocol.

We also estimate human reference performance. To obtain a manageable reference estimate, we randomly sampled 700 VQA questions from the full 193K MCVQA set. We recruited six undergraduate volunteers who expressed interest in the project after it was introduced by teaching assistants during regular class sessions. Participation was voluntary, and each participant spent approximately 1.5–2 hours completing the evaluation in a provided laboratory room. No monetary payment or other compensation was provided; therefore, an assessment of whether payment was adequate for participants' demographics (e.g., country of residence) is not applicable. Three participants completed the 700-question evaluation without additional guidance, while the other three were instructed to pay particular attention to common fire-and-smoke ambiguities, such as distinguishing smoke from water vapor, fog, or haze, and distinguishing flames from streetlights or electric arcs.

As shown in Table 12, the non-Qwen-family models achieve an average accuracy of 60.6% on the full SAFIRE MCVQA benchmark, while the Qwen family achieves an average accuracy of 64.6%, with the best Qwen-family model reaching 65.8%. This suggests that a generator-family effect may exist, but its magnitude is limited compared with the gap between current MLLMs and human performance. The human study is conducted on a randomly sampled 700-question subset as a reference estimate rather than a strict paired full-set comparison. Nevertheless, the average human accuracy of 82.5% is substantially higher than both the non-Qwen average and the best Qwen result. This indicates that the human–model gap is much larger than the gap among model families, suggesting that SAFIRE reflects fire-smoke reasoning difficulty for current MLLMs rather than merely generator-family bias.

Table 12: **Generator-family bias analysis and human reference performance.** Models are evaluated on the full MCVQA set, human performance* is evaluated on a randomly sampled 700 MCVQA subset.

Model group	Model / Mode	Acc.
Non-Qwen family	Step3-VL-10B	63.4%
	InternVL3.5-38B	62.7%
	LLaVA-OneVision-2-8B	61.6%
	GLM-4.1V-9B-Base	59.7%
	Gemma-4-E4B-it	59.3%
	Molmo2-8B	56.9%
	Average	**60.6%**
Qwen family	Qwen3.6-27B non-thinking	65.8%
	Qwen3.6-35B-A3B non-thinking	65.5%
	Qwen3.5-9B non-thinking	64.9%
	Qwen3.5-9B thinking	62.2%
	Average	**64.6%**
Human performance*	Unguided	81.3%
	Guided	83.6%
	Average	**82.5%**

J MLLM Supervised Fine-Tuning Study

The main focus of SAFIRE is benchmark construction and evaluation. Nevertheless, we conduct an additional supervised fine-tuning experiment to examine whether SAFIRE can also provide useful training supervision.

We fine-tune Qwen3-VL-8B-Instruct with LLaMA-Factory using LoRA rather than full fine-tuning. The vision encoder is frozen, and LoRA adapters are applied to the language-model projection and feed-forward layers (`q/k/v/o_proj` and `gate/up/down_proj`; rank 16, alpha 32, dropout 0.05). Training uses the `qwen3_vl_nothink` template with maximum sequence length 1,024 and maximum image pixels 262,144. We train for one epoch using an effective batch size of 32, learning rate 1e-4 with cosine scheduling and 126 warmup steps, weight decay 0.01, random seed 42, and evaluate the final checkpoint.

To avoid information leakage, we split the data at the image level rather than at the question level, ensuring that MCVQAs generated from the same image are not assigned to different splits. The

Table 13: **Supervised fine-tuning result on SAFIRE.** Qwen3-SAFIRE-SFT is initialized from Qwen3-VL-8B and fine-tuned on the image-level training split.

Model	Acc.	Correct / Total
Qwen3-VL-8B Baseline	63.02%	12,189 / 19,340
Qwen3-SAFIRE-SFT	70.83%	13,699 / 19,340
Improvement	+7.81 pp	+1,510 correct

full 193K MCVQA set is split into 70% training data, 20% reserved data, and 10% held-out test data. This gives approximately 135.1K MCVQA samples for supervised fine-tuning and 19,340 MCVQA samples for testing.

As shown in Table 13, supervised fine-tuning on SAFIRE substantially improves Qwen3-VL-8B, increasing held-out accuracy from 63.02% to 70.83% (+7.81 percentage points). In terms of errors, the baseline makes 7,151 mistakes on the held-out set, while Qwen3-SAFIRE-SFT makes 5,641 mistakes, corresponding to a 21.12% relative reduction in error rate. The improvement is also broad rather than concentrated in a small subset: all 20 scenarios improve, with gains ranging from +5.07 to +10.26 percentage points, and all 10 reasoning dimensions improve, with gains ranging from +1.91 to +16.29 percentage points. These results indicate that SAFIRE is not only useful for evaluating existing MLLMs, but also provides effective supervision for improving fire-smoke reasoning through task-specific fine-tuning.

K Visual Diversity Across Fire-Smoke Scenarios

Fig. 10 illustrates ensemble examples from SAFIRE's 20 distinct fire and smoke scenarios. While Figures 11, 12, and 13 showcase 80 additional representative images from the SAFIRE benchmark, covering all 20 fire-smoke scenarios. These visualizations further highlight the dataset's diversity in fire behavior, smoke characteristics, and contextual settings.

Each scenario includes 4-5 illustrative samples to capture intra-class variation (e.g., perspective, environment, object presence) and inter-class distinctiveness. For example, **Fig. 11** presents large-scale natural fires such as *Volcano* and dynamic man-made events like *Aerospace Operations* and *Burning Vehicles*. **Fig. 12** features culturally embedded and industrial scenarios including *Torch Fire*, *Metal Forging Fire*, and *Explosion*, each with unique flame structures and textures. **Fig. 13** covers smaller or ambiguous sources such as *Smoking*, *Gas Stove Fire*, and *Meteorite Fire*, which are essential for broadening the benchmark beyond conventional scene types.

Figure 9: Example image from the Volcano scenario.

The scenarios align with the dataset folder structure on our Hugging Face repository, supporting reproducibility and traceability. These extended visual samples emphasize SAFIRE's focus on realistic, fine-grained fire-smoke phenomena across natural, industrial, accidental, civil controlled, and recreational contexts, and help illustrate the generalization challenges faced by both vision-only and multimodal models.

L A Full Generated MCVQA Example

Example Annotation Overview. The image `volca_10.jpg` in Fig.9 depicts a dramatic volcanic eruption occurring at night.The comprehensive metadata includes a set of 20 multiple-choice visual questions that test various levels of multimodal reasoning. These questions assess quantitative perception (one distinct fire/smoke source), and physical properties (flame color, smoke density and color). Higher-order reasoning questions probe the likely cause of the fire (natural eruption), the absence of human or animal life, and the expected public reaction (evacuation). Moreover, abstract linguistic reasoning is also evaluated, such as interpreting idiomatic expressions like "on fire" and predicting plausible future developments (e.g., lava activity will continue). Details can be found in Figures 14, and 15. Overall, we pick out this example here to illustrate how SAFIRE's MCVQA framework captures both factual visual grounding and inferential logic, providing a rigorous benchmark for assessing the safety-relevant capabilities of modern multimodal models.

Figure 10: **Illustration of SAFIRE's 20 diverse fire and smoke scenarios**, designed to capture both common and uncommon events across natural, civil, industrial, and symbolic contexts. From top-left to bottom-right: (1) Open Grassland Fire, (2) Forest Fire, (3) Volcano, (4) Meteor, (5) Aerospace Operations, (6) Flare Stack, (7) Metal Forging, (8) Residential Scene, (9) Vehicle Fire, (10) Waste Disposal Scene, (11) Explosion, (12) Firework, (13) Hot Air Balloon, (14) Barbecue, (15) Campfire-Bonfire, (16) Torch Parade, (17) Candlelight, (18) Smoking, (19) Incense Burning, (20) Gas Stove.

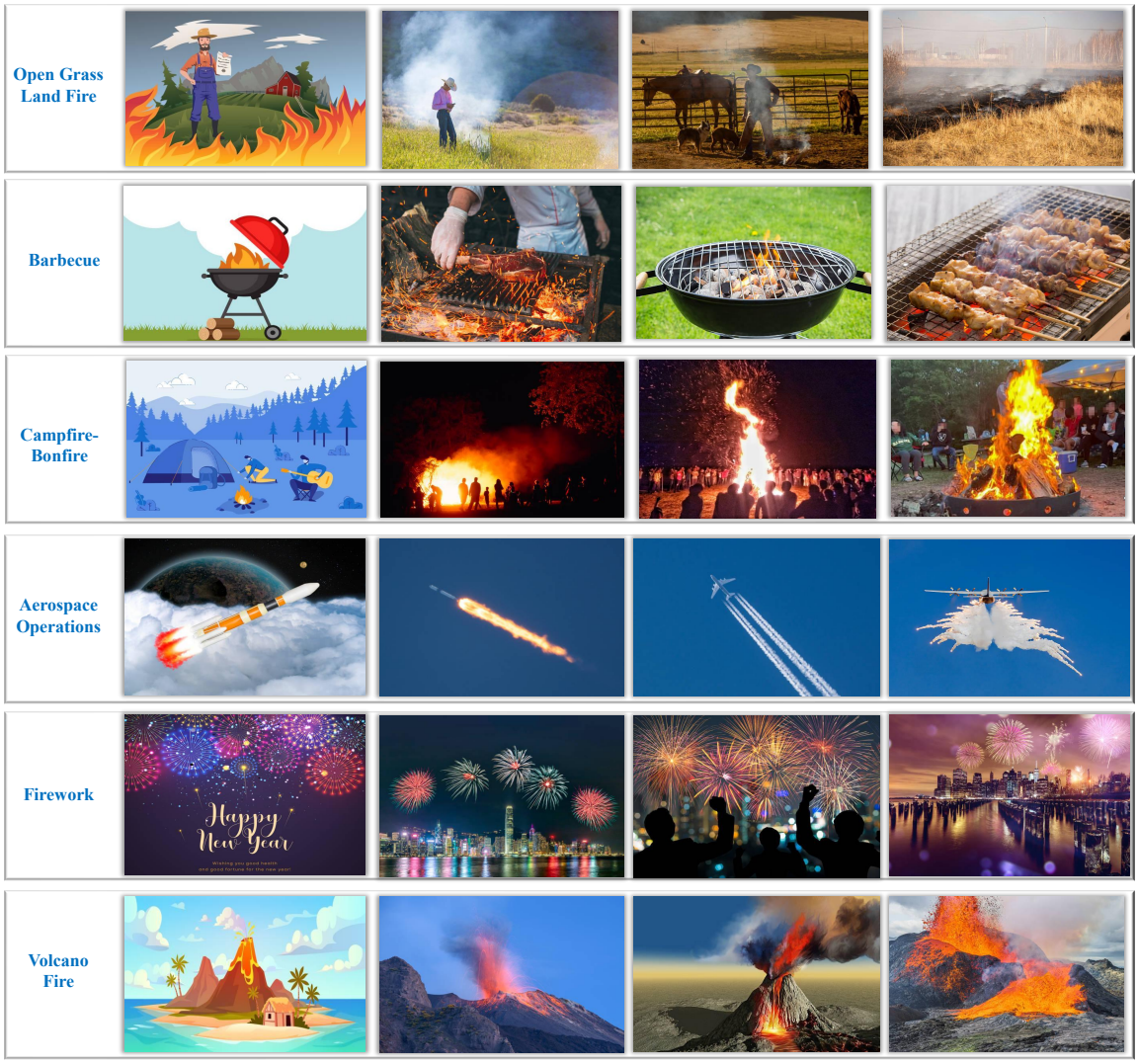

Figure 11: Visual examples of fire-smoke scenarios used in the benchmark.

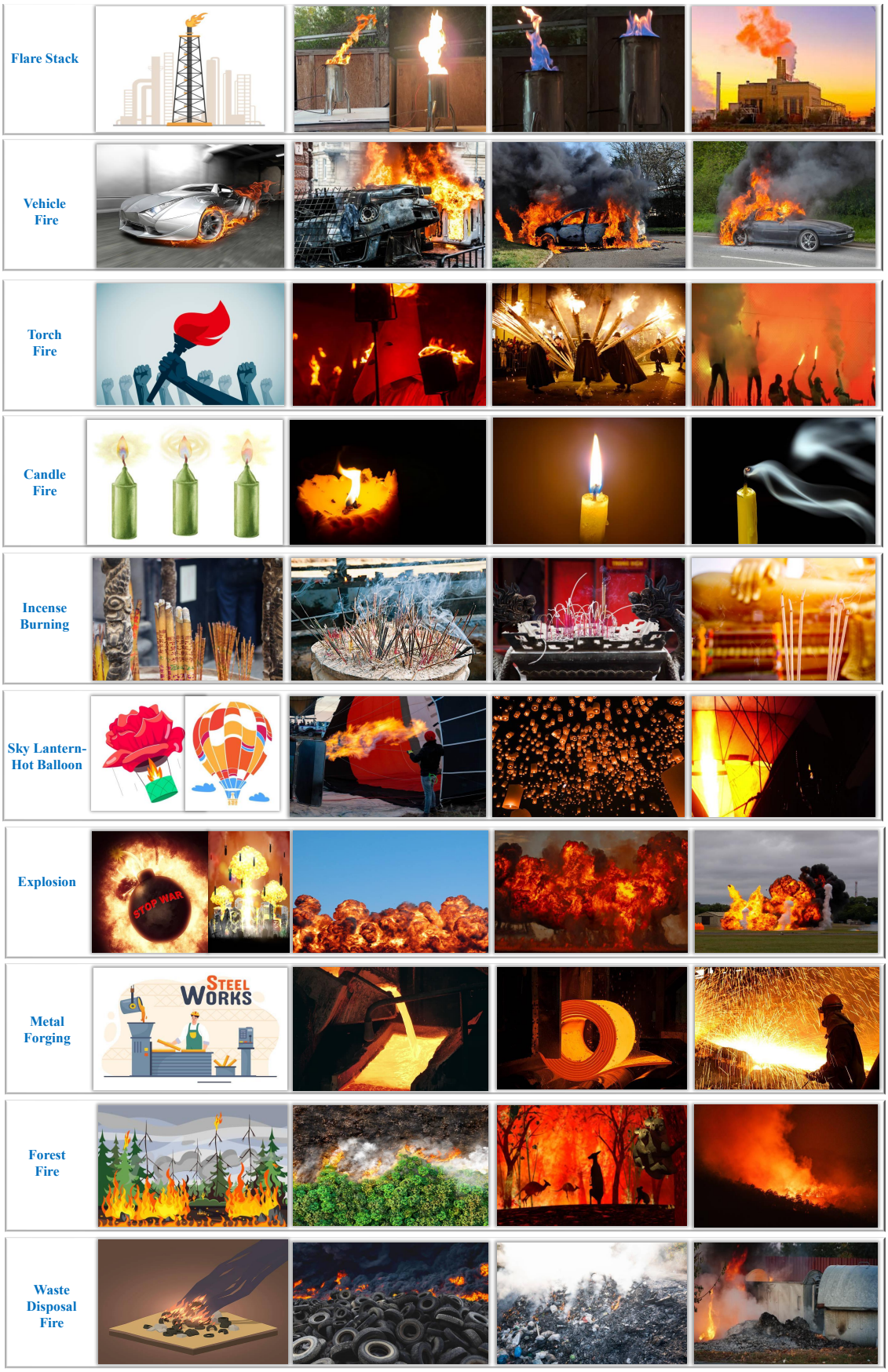

Figure 12: Visual examples of fire-smoke scenarios used in the benchmark.

Figure 13: Visual examples of fire-smoke scenarios used in the benchmark.

MCVQA JSON Annotation (Part 1)


```
{
  "Image_name": "volca_10.jpg",
  "QA_pairs": [
    {
      "question": "What kind of smoke/flame is this?",
      "options": ["A. Bright lava eruption at night", "B. Geothermal steam vent", "C. Building
           fire", "D. Fire icon"],
      "answer": "A. Bright lava eruption at night"
    },
    {
      "question": "Where is the fire located in the scene?",
      "options": ["A. Left side", "B. Right side", "C. Center", "D. Not visible"],
      "answer": "C. Center"
    },
    {
      "question": "How many distinct fire or smoke sources are visible?",
      "options": ["A. None", "B. One", "C. Two", "D. More than two"],
      "answer": "B. One"
    },
    {
      "question": "What is the apparent danger level of this fire?",
      "options": ["A. Low", "B. Moderate", "C. High", "D. Extreme"],
      "answer": "D. Extreme"
    },
    {
      "question": "How much smoke is visible in the image?",
      "options": ["A. No smoke", "B. Light", "C. Moderate", "D. Thick smoke"],
      "answer": "D. Thick smoke"
    },
    {
      "question": "What is the color of the smoke?",
      "options": ["A. White", "B. Black", "C. Gray", "D. No smoke"],
      "answer": "C. Gray"
    },
    {
      "question": "Are flames visible?",
      "options": ["A. Yes", "B. No", "C. Not sure", "D. Not applicable"],
      "answer": "A. Yes"
    },
    {
      "question": "What is the color of the flames?",
      "options": ["A. Orange/Yellow", "B. Red", "C. Blue/White", "D. No flames"],
      "answer": "A. Orange/Yellow"
    },
     {
      "question": "What object or material is burning?",
      "options": ["A. Lava or volcanic material", "B. Vegetation", "C. Nothing is visibly
          burning", "D. Not clear"],
      "answer": "A. Lava or volcanic material"
    },
     {
      "question": "Is this a real fire scene or just an icon or sketch?",
      "options": ["A. Real fire", "B. Icon/Symbol/sketch", "C. Simulation", "D. Not sure"],
      "answer": "A. Real fire"
    },
    {
      "question": "What likely caused the situation?",
      "options": ["A. Natural volcanic eruption", "B. Graphic design", "C. Industrial process"
          , "D. Unknown"],
      "answer": "A. Natural volcanic eruption"
    },
```


Figure 14: A generated MCVQA annotation example from the Volcano scenario, Part 1.

MCVQA JSON (Part 2)


```
    {
      "question": "Are there people or animals in the image?",
      "options": ["A. No", "B. People present", "C. Animals present", "D. Not sure"],
      "answer": "A. No"
    },
    {
      "question": "Is the fire being actively controlled?",
      "options": ["A. Yes", "B. No", "C. Not applicable", "D. Cannot tell"],
      "answer": "B. No"
    },
    {
      "question": "What is the size of the fire?",
      "options": ["A. Small", "B. Medium", "C. Large", "D. Massive"],
      "answer": "D. Massive"
    },
    {
      "question": "What time of day does the image suggest?",
      "options": ["A. Daytime", "B. Nighttime", "C. Indoor lighting", "D. Not determinable"],
      "answer": "B. Nighttime"
    },
    {
      "question": "What is the spatial relationship between the smoke and the fire?",
      "options": ["A. Smoke above the fire", "B. Smoke to the left of Fire", "C. Smoke to the
          right of Fire", "D. Not applicable"],
      "answer": "A. Smoke above the fire"
    },
    {
      "question": "What will happen next to the Fire/Smoke?",
      "options": ["A. Eruption or lava activity continues", "B. Smoke disperses", "C. Fire
          quenches", "D. Explosion occurs"],
      "answer": "A. Eruption or lava activity continues"
    },
    {
      "question": "What is the purpose of this fire?",
      "options": ["A. Natural volcanic eruption", "B. Industrial operation", "C. Celebration",
           "D. Ritual use"],
      "answer": "A. Natural volcanic eruption"
    },
    {
      "question": "How would people react to this smoke/fire?",
      "options": ["A. Take photos only", "B. Ignore it", "C. Evacuate the area", "D. Watch
          from nearby"],
      "answer": "C. Evacuate the area"
    },
    {
      "question": "What does 'on fire' mean here?",
      "options": ["A. Literal fire", "B. Excited situation", "C. Under attack", "D. Figurative
           speech"],
      "answer": "A. Literal fire"
    }
  ]
}
```


Figure 15: A generated MCVQA annotation example from the Volcano scenario, Part 2.