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

Anonymous ACL submission

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 existing benchmarks lack diverse real-world scenarios and context-aware evaluation. We introduce SAFIRE, the first large-scale benchmark for fire-smoke understanding in MLLMs, comprising 105K images from 20 scenarios and 224K multiple-choice VQA (MCVQA) pairs across 10 evaluation dimensions, from basic perception to higher-order reasoning. A multi-stage filtering pipeline with model-guided majority voting ensures annotation quality. Evaluating seven open-source MLLMs (7B-72B) yields 57.5% average accuracy, exposing major gaps in safety-critical reasoning. We further provide a domain-specific image-caption set and show that adapting vision encoders with only 3% of our supervised data boosts fire-scene classification by 12.7%, underscoring the inadequacy of generic features and the potential of lightweight adaptation. Our code, datasets, evaluation pipeline, and pretrained models will be released publicly.

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).

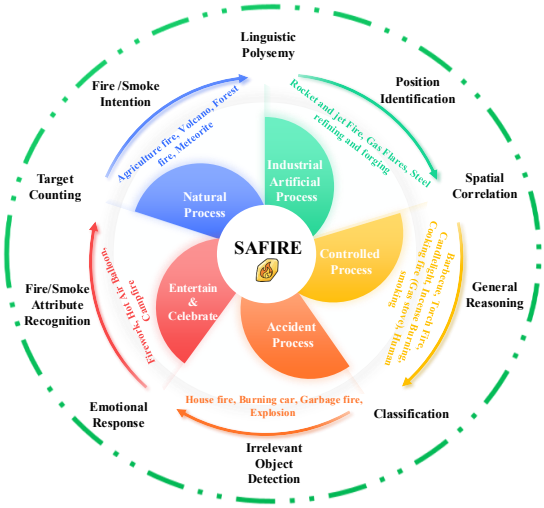

Figure 1: **From windmill center to the periphery:** SAFIRE consists of 5 processes spanning 20 real-world scenarios, with 10 evaluation dimensions and over 224K curated question-answer pairs derived from our subsets.

Recent MLLMs demonstrate strong zero-shot performance and richer contextual reasoning than traditional models (Liu et al., 2023a; Bai et al., 2025), yet existing fire–smoke benchmarks fail to evaluate the situational reasoning required in safety-critical contexts (Tab. 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.

(a) Misleading “White Smoke" (b) Awaring “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), the first large-scale benchmark for comprehensive, context-aware evaluation of MLLMs in fire–smoke reasoning. As shown in Fig. 1, SAFIRE comprises 105K curated images from 20 real-world scenarios and 224K multiple-choice VQA pairs spanning 10 evaluation dimensions, supporting both perceptual and higher-order reasoning at the scene level. Our main contributions are summarized as follows:

- To the best of our knowledge, *SAFIRE* is the first multi-scenario benchmark explicitly designed to evaluate MLLMs' reasoning in fire–smoke contexts. Annotation quality is ensured through multi-stage filtering and model-guided majority voting. *Code and data will be released publicly.*
- *SAFIRE* enables systematic assessment of scene-level, contextual, causal, and behavioral reasoning, providing a testbed that extends beyond conventional classification or detection tasks.
- Evaluations of seven open-source MLLMs (7B–72B) shows 57.5% average MCVQA accuracy; CLIP-style encoders achieve up to 62.7% in zero-shot scene classification. Adapting the vision encoder with 3% supervision improves accuracy by 12.7%, highlighting remaining gaps and the value of lightweight domain adaptation.

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 *MAFire-Net* (Khan et al., 2023) 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 emphasized localization and multi-scenario annotations. Datasets such as *FiSmo* (Cazzolato et al., 2017), *Nemo* (Yazdi et al., 2022), *DFS* (Wu et al., 2023), and *HQFSD* (Qian et al., 2024) incorporate bounding boxes, severity stages, or “fire-like” distractor classes to reduce false positives. Advances in object detection models from YOLO (Redmon et al., 2016) to DETR (Carion et al., 2020) and DINO (Zhang et al., 2022) have further improved scene localization. Segmentation and sensor-augmented datasets such as *WildfireSpread* (Gerard et al., 2023), *MultiFire20K* (Han et al., 2024a), and *UniInd-FireSmoke* (Li et al., 2025b) expand modalities via satellite or industrial sensors, yet they primarily address perception rather than multimodal reasoning.

Table 1: **Comparison of SAFIRE with Previous Influential Image-based Works.** Here, "Influential works" refer to either widely cited historical benchmarks (2015–2020) or foundational recent benchmarks (2021–2025). denotes Classification, denotes Detection, denotes Segmentation, denotes Sensor Data, denotes MCVQA and Captions. The proposed SAFIRE benchmark stands out for its overall scale, scenario diversity, and comprehensive coverage.

Benchmark	Anno.	Fire	Smoke	# Categories	# 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
MAFire-Net (2022)		✓	✗	2	18,881
EdgeFireSmoke (2022)		✓	✓	8	86,775
UFS-data (2022)		✓	✓	8	849,640
MS-FSDB (2024)		✓	✓	35	12,586
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
Mlich (2020)		✓	✗	1	6,386
WildfireSpread(2023)		✓	✗	2	13,607
MultiFire20K (2024)		✓	✓	4	20,500
UniInd-FireSmoke(2025)		✓	✓	6	30,101
SAFIRE (Ours)		✓	✓	20	105,985

Landscape summary. Table 1 summarizes 19 representative fire–smoke datasets (2015–2025). Classification tasks dominate (15/19), with fewer detection or segmentation resources. None include image–caption or VQA pairs for evaluating reasoning, and over half 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), *UFS-data* (Hosseini et al., 2022), and *EdgeFireSmoke* (Almeida et al., 2022)—offer relatively diverse scene coverage. Misuse of terms such as "detection" for classification further obscures annotation depth.

Rise of multimodal models. Recent advances in vision–language modeling, from encoder-only *CLIP* (Radford et al., 2021) to multimodal LLMs such as *LLaVA* (Liu et al., 2023a), *LLaMA 3.2 Vision* (Meta AI, 2024), *Qwen-VL* (Bai et al., 2025), and *Gemma-3* (Gemma Team, 2025), enable zero-/few-shot transfer and contextual reasoning. However, no prior benchmark systematically evaluates safety-critical *reasoning* for fire–smoke understanding at scale. **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.

Table 2: **SAFIRE dataset statistics across 20 fire–smoke scenarios.** QA-I indicates the image subset used for generating questions, and QA-P represents the total number of generated question–answer pairs.

Scenario	Full Set	QA-I (%)	QA-P
Natural Process			
Agricultural	6,789	777 (11.4%)	15,500
Forest	5,751	651 (11.3%)	12,917
Volcano	6,365	805 (12.6%)	16,099
Meteorite	5,936	460 (07.7%)	9,089
Industrial Process			
Rocket Fire	4,139	466 (11.3%)	9,314
Gas Flares	2,356	465 (19.7%)	9,133
Steel Forge	3,163	320 (10.1%)	6,443
Accident Process			
House Fire	4,681	490 (10.5%)	9,614
Car Fire	2,646	288 (10.9%)	5,759
Garbage Fire	5,902	686 (11.6%)	13,719
Explosion	6,624	786 (11.9%)	15,682
Entertainment			
Firework	6,605	760 (11.5%)	15,155
Hot Balloon	1,810	209 (11.5%)	4,180
Barbecue	6,787	780 (11.5%)	15,600
Campfire	6,160	705 (11.4%)	14,100
Controlled Process			
Torch Fire	5,822	704 (12.1%)	14,076
Candlelight	5,943	702 (11.8%)	14,040
Smoking	5,922	349 (05.9%)	6,912
Incense Burning	6,783	791 (11.7%)	15,792
Gas Cooking	5,775	697 (12.1%)	13,843
Total	**105,959**	**11,891 (11.2%)**	**224,050**

3 SAFIRE Benchmark

We introduce **SAFIRE**, a large-scale benchmark for evaluating context-aware fire–smoke understanding in MLLMs. It contains 105K images across 20 real-world scenarios and a curated subset of 11.9K images with 224K 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.

Data Collection and Structuring. Images and captions are gathered from multiple sources and filtered through a multi-factor process combining automatic keyword matching, semantic validation, and manual review, yielding 105K high-quality samples (Fig. 4, Stage I). From this corpus, we select 10.9K samples to build two subsets under a unified taxonomy: (i) an MCVQA set spanning ten reasoning dimensions and (ii) an enhanced image–caption set covering five fire-type categories

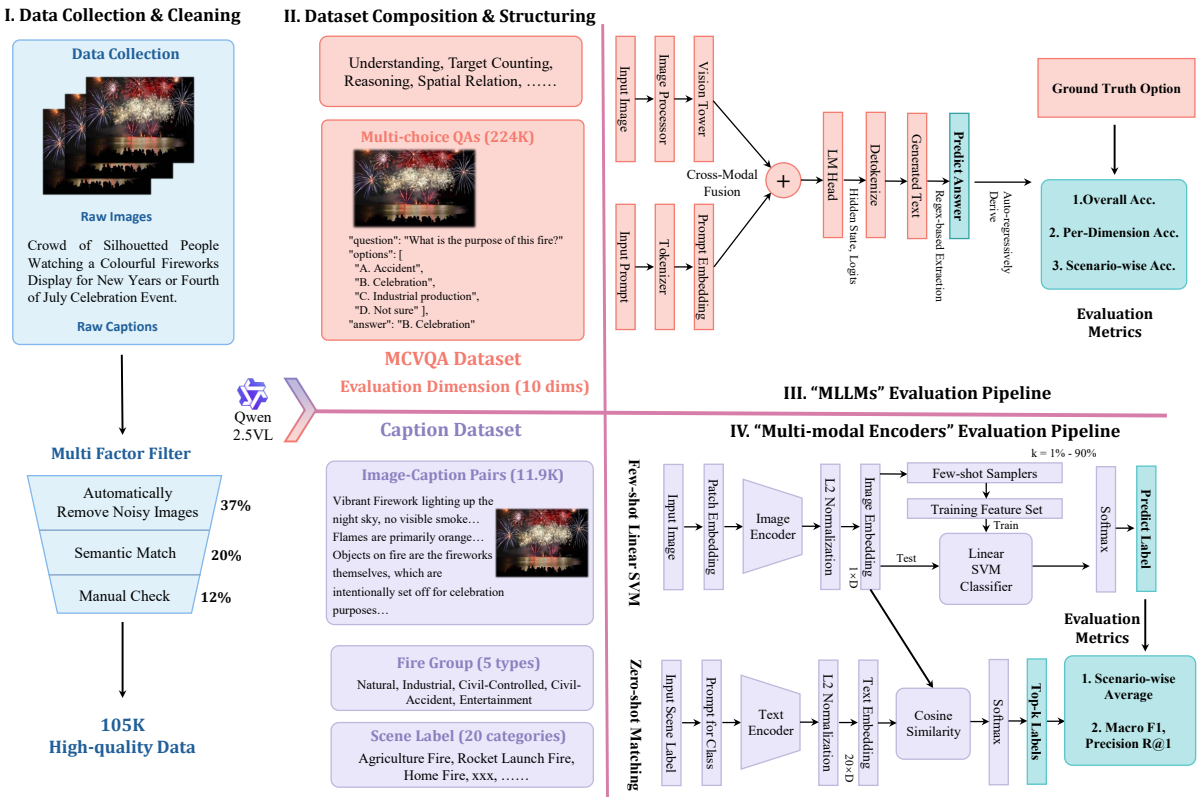

Figure 4: **SAFIRE pipeline.** Four stages: (I) multi-source collection and filtering; (II) construction of MCVQA and caption branches with hierarchical labels; (III) MLLM evaluation across ten reasoning dimensions; (IV) benchmarking of vision–language encoders under zero-shot and few-shot settings.

and twenty scene labels (Fig. 4, Stage II).

Annotation Pipeline. All annotations are generated using a structured prompting framework built on Qwen2.5-VL-72B-Instruct (Bai et al., 2025). 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.

Quality Assurance. Annotation quality is ensured through multi-stage validation. We employ majority voting across multiple MLLMs (Fig. 8) to flag inconsistent samples, which are then manually reviewed and corrected. This combination of automated checks and human verification provides scalable and reliable quality control.

Reasoning Dimensions. The ten reasoning dimensions are *target counting, classification, general reasoning, fire/smoke intention, emotional response, linguistic polysemy, attribute recognition, spatial correlation, position identification*, and *irrelevant object detection*, covering perception, spatial grounding, and causal understanding. Each image in the MCVQA set is paired with 20 multiple-choice questions distributed across four tiers: (1) *Scene-level* (scenario classification), (2) *Spatial and instance-level* (position, number of sources, danger level), (3) *Recognition* (smoke color, flame visibility, lighting context), and (4) *Reasoning and context* (source, human activity, trend prediction, emergency response).

Taxonomy and Dataset Statistics. The five fire-type categories are *natural, industrial, accidental, controlled*, and *celebratory*, further divided into twenty real-world scenarios such as *forest, steel forge, car, fireworks*, and *incense burning*. Table 2 summarizes per-scenario counts of images and question–answer pairs.

Visualization. Figure 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.

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

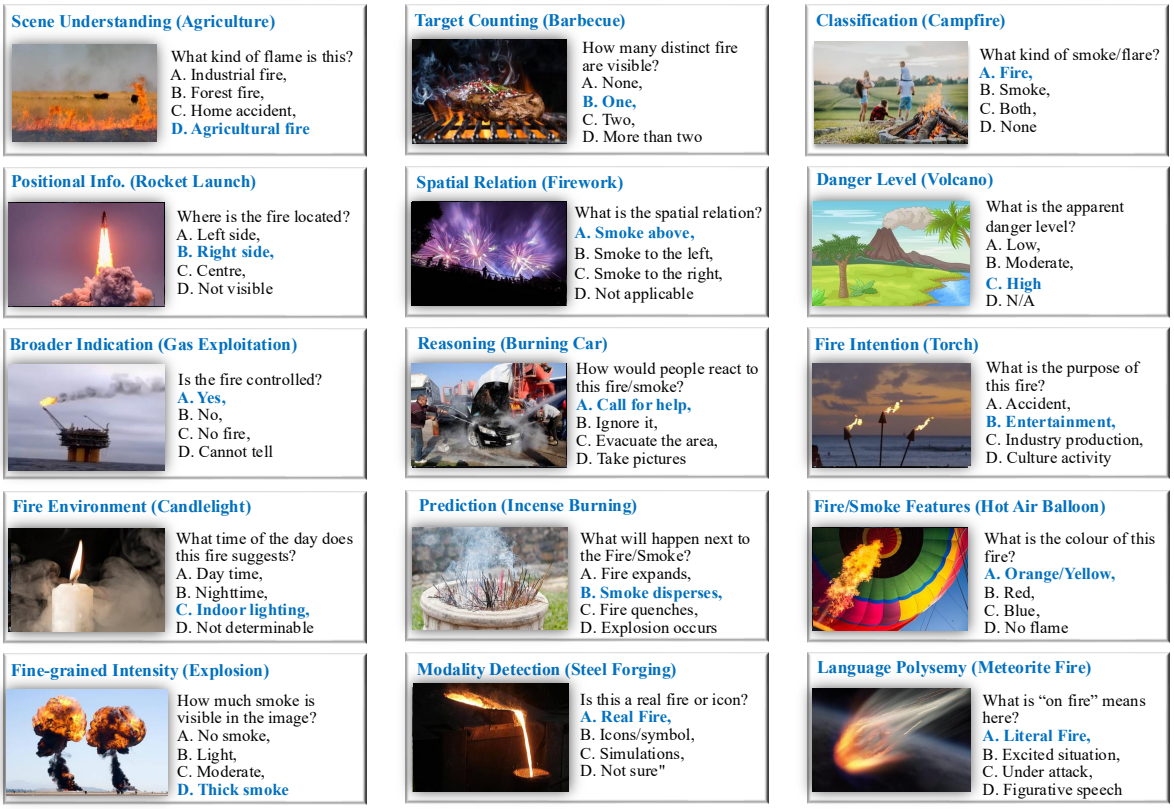

Figure 5: **Illustrative MCVQA samples.** Each question type (for example, scene understanding or target counting) is aligned with images from distinct fire–smoke scenarios.

the 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 D.

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, Stage IV).

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

Few-shot. A linear SVM is trained on frozen image embeddings using small supervision ratios 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 D.

4 Experiments

4.1 Experimental Setup

We evaluate both MLLMs and vision–language encoders on the SAFIRE benchmark, as illustrated in Figure 4. Stage III covers MLLM inference on the MCVQA task, while Stage IV evaluates encoder performance through the FireCLIP variant.

MLLM Configuration. All MLLM experiments were conducted on an NVIDIA A100 multi-GPU cluster using the vLLM inference engine with automatic device mapping and `bfloat16` precision. We evaluated seven instruction-tuned vision–language models: LLaMA-3.2-Vision (11B) (Meta AI, 2024), *Qwen-2.5-VL (7B, 32B, 72B)* (Bai et al., 2025), *Mistral-3.1 (24B)* (Mistral AI, 2025), and *Gemma-3 (12B, 27B)* (Gemma Team, 2025). 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.

FireCLIP and Encoder Config. For Stage IV, we evaluate vision–language encoders using both pretrained and domain-adapted variants. We fine-

Table 3: **Accuracy (%) of MLLMs on the 20-scenario fire–smoke MCVQA benchmark. Bold** indicates the highest accuracy in each column (i.e., best-performing per scenario), and underlined indicates the lowest.

Model	Size	Natural Process				Industrial Process			Accident Process			
	(B)	Agricul.	Volcano	Forest	Meteorite	Rocket	Indust.	Steel	Explos.	House	Car	Garbage
Qwen2.5-VL	7	50.3	62.7	63.1	59.2	50.1	**59.8**	59.5	57.0	61.7	49.2	58.4
LLaMA-3.2	11	46.0	47.6	54.4	37.7	43.3	52.1	42.7	41.8	49.1	46.0	42.7
Gemma-3	12	50.3	61.9	62.0	56.1	51.3	57.8	57.4	53.6	61.7	48.9	58.5
Mistral-3.1	24	49.7	64.4	59.4	58.2	50.2	55.1	59.9	58.0	61.7	49.3	58.9
Gemma-3	27	**54.5**	61.7	**66.0**	54.4	51.9	56.0	55.6	56.3	63.2	48.9	**61.3**
Qwen2.5-VL	32	49.5	64.4	62.4	60.1	51.6	58.6	57.3	59.3	63.4	**51.3**	61.1
Qwen2.5-VL	72	53.0	**67.2**	65.8	**62.8**	**53.2**	59.7	**62.5**	**59.5**	**66.5**	50.8	61.1
Scene Avg.	–	50.5	61.4	61.9	55.5	50.2	57.0	56.4	55.1	61.0	49.2	57.4

Model	Size	Entertain & Celebrate				Civil-Controlled Process					All
	(B)	Firework	Hot Balloon	Barbecue	Campfire	Torch	Candle	Incense	Cooking	Smoking	Avg.
Qwen2.5-VL	7	60.8	59.2	54.9	**62.4**	61.9	55.5	62.1	65.1	60.5	58.6
LLaMA-3.2	11	43.4	47.8	45.6	47.0	43.4	41.7	47.6	44.8	37.8	44.8
Gemma-3	12	**66.4**	63.8	54.7	59.8	**65.6**	55.9	62.5	61.1	62.1	58.6
Mistral-3.1	24	64.4	59.5	56.3	57.0	60.9	60.9	63.8	68.2	64.5	59.2
Gemma-3	27	64.9	61.4	55.2	58.2	63.8	53.7	58.6	62.2	64.1	58.7
Qwen2.5-VL	32	62.5	56.3	56.8	57.4	61.5	59.3	66.4	**70.5**	**65.7**	59.8
Qwen2.5-VL	72	64.6	**63.5**	**60.1**	61.4	65.9	**63.6**	**68.7**	**73.1**	68.1	**62.7**
Scene Avg.	–	61.0	58.8	54.8	57.6	60.4	55.8	61.4	**63.6**	60.4	57.5

tune CLIP (ViT-L/14 backbone) on the SAFIRE image–caption subset to obtain *FireCLIP*, using 93K images with their raw captions for training and 11.9K curated subset for testing on two A100 GPUs with the official CLIP implementation. FireCLIP is compared against pretrained baselines, including *EvaCLIP* (Liu et al., 2023b) and *OpenCLIP* (Cherti et al., 2023), to assess domain adaptation and generalization in fire–smoke understanding.

4.2 MLLM Evaluation Results

We evaluate MLLMs in **Stage III** of the SAFIRE benchmark to assess their ability to solve multimodal multiple-choice questions (MCVQA). This evaluation measures model performance across ten reasoning dimensions and twenty real-world scenarios 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) show that most state-of-the-art vision–language models achieve modest performance, with average accuracies below 63%. 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, Qwen2.5-VL 72B outperforms its 32B counterpart by only three points, suggesting that scaling alone is insufficient without domain adaptation. Data quality and task relevance appear more influential than size, as Qwen2.5-VL consistently surpasses larger models such as Gemma-3 and Mistral-3. Performance also varies widely by scenario: industrial and emergency settings (e.g., car fires, rocket launches) yield the lowest accuracies, whereas controlled or familiar scenes (e.g., cooking, forest, and volcanic smoke) perform significantly better. This disparity likely stems from both data imbalance and limited exposure to rare industrial 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 *Fire/Smoke Spatial Correlation* reach the highest accuracies (73.4% and 83.1%), showing that MLLMs handle global scene recognition and spatial relationships well. In contrast, reasoning-heavy tasks like *Emotional Response* (38.0%) and *Fire/Smoke Intention* (51.7%) remain challenging, highlighting persistent difficulty in understanding causality and human-centered semantics. Spatial grounding also lags behind: *Position Identification* averages only 50.7%, indicating that models strug-

Table 4: **Accuracy (%) across 10 evaluation dimensions on SAFIRE. Bold** indicates the highest accuracy per row (i.e., best-performing model for each dimension), while underlined indicates the lowest accuracy.

Dimension	**Qwen2.5VL 7B**	**LLaMA3.2 11B**	**Gemma3 12B**	**Mistral3.1 24B**	**Gemma3 27B**	**Qwen2.5VL 32B**	**Qwen2.5VL 72B**	**Avg. (%)**
Target Counting	**63.4**	55.8	49.0	62.3	51.2	59.8	60.3	57.4
Classification	72.4	73.3	73.2	73.6	**75.7**	72.6	73.3	73.4
General Reasoning	56.3	30.9	59.1	57.7	55.0	57.7	**63.3**	54.3
Fire/Smoke Intention	49.9	54.3	51.3	47.2	**56.1**	50.8	52.0	51.7
Emotional Response	36.0	21.4	34.9	34.0	37.8	44.7	**57.1**	38.0
Linguistic Polysemy	54.4	20.8	65.1	74.3	71.8	71.2	**78.3**	62.3
Fire/Smoke Attributes	53.7	49.3	49.3	50.5	49.9	52.7	**54.9**	51.5
Fire/Smoke Spatial Corr.	84.7	80.7	84.7	82.2	83.4	78.2	**87.8**	**83.1**
Position Identification	51.7	35.4	**63.1**	49.6	56.9	49.3	48.7	50.7
Irrelevant Object Detect.	67.4	31.5	59.5	64.5	**75.3**	74.0	61.6	62.0
Model Average	59.0	45.3	58.9	59.6	61.3	61.1	**63.7**	58.4

gle to localize fires under occlusion or subtle conditions. Interestingly, reasoning performance does not scale linearly with model size; smaller models such as Qwen2.5-VL 7B outperform larger ones like Mistral-3.1 24B on several reasoning tasks, underscoring the impact of training corpus quality over scale. Tasks involving semantic ambiguity, such as *Linguistic Polysemy*, show the widest gap (20.8%–78.3%), confirming that context disambiguation remains a major weakness. By contrast, visually grounded or binary tasks like *Irrelevant Object Detection* and *Fire/Smoke Attributes* achieve stable mid-to-high performance, benefiting from clear visual cues rather than deep reasoning. 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.

Summary and Implications. Across scenarios and reasoning dimensions, current MLLMs demonstrate strong scene recognition and spatial correlation but remain limited in affective interpretation, causal reasoning, and fine-grained localization. Performance gaps across models reflect the combined influence of scale, architecture, and data quality. SAFIRE's rich annotations and hierarchical design facilitate detailed analysis of these limitations and offer guidance for future work, including domain-specific pretraining, targeted data augmentation for rare or ambiguous scenes, and improved integration of visual and textual reasoning. Overall, SAFIRE provides a comprehensive foundation for building reliable, context-aware multimodal systems for safety-critical fire–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 analysis examines how well vision–language encoders generalize to diverse fire–smoke scenarios under varying levels of label supervision. Experiments span five zero-shot models and five supervised variants across twenty scenarios (Fig. 6, 7).

Scenario-level Analysis. Figure 6 visualizes scenario-wise accuracies. Zero-shot variants such as `FireCLIP_ViT_L14` and `EvaCLIP` (Liu et al., 2023b) substantially outperform the 5% random baseline. Highly salient categories such as *Human Smoking*, *Steel Refining Fire*, and *Volcano* exceed 90% accuracy, while visually cluttered or ambiguous scenes such as *Garbage Fire*, *Candlelight*, and *Torch Parade* show weaker results. Model-wise variation is also evident: *Meteorite* achieves over 70% accuracy in `EvaCLIP` but only 19% in `CLIP_ViT_B32`, suggesting reliance on global scene cues. Scenarios such as *Campfire* and *Barbecue* benefit from web-scale corpora used by models like `OpenCLIP` (Cherti et al., 2023) and `EvaCLIP`, which likely contain related imagery.

Metric-level Analysis. Figure 7 compares macro-level performance across ten encoders. Among zero-shot models, `FireCLIP_ViT_L14` achieves the best balance (F1=0.64, Recall@1=0.65, Precision=0.64), demonstrating the benefits of domain-specific pretraining.

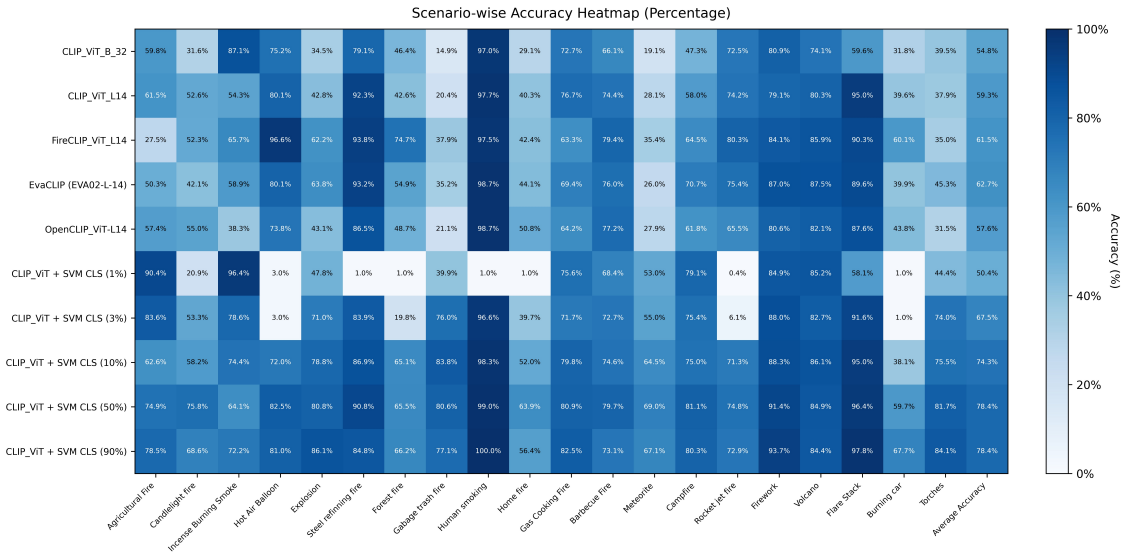

Figure 6: **Zero-shot accuracy of five multi-modal encoders** (i.e., CLIP-based models) across 20 scenarios in the SAFIRE benchmark. The last column shows results for CLIP+SVM, which adds a simple tunable classification head. A random guess across 20 classes yields 5% accuracy.

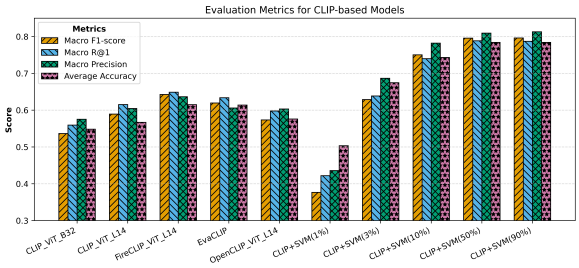

Figure 7: **Macro performance metrics (F1, Recall@1, and Precision) across ten CLIP-based models.** The first five are zero-shot variants, while the latter are CLIP+SVM models with increasing label supervision.

Generic models such as `CLIP_ViT_B32` and `OpenCLIP_ViT_L14` lag behind (F1=0.54–0.57), underscoring the limits of domain-agnostic embeddings. Models like `EvaCLIP` exhibit higher recall than precision, indicating broader generalization but increased false positives. Supervised variants (CLIP+SVM) show consistent gains as label supervision increases. With 90% supervision, CLIP+SVM achieves F1=0.80 and Precision=0.81, while even 3% supervision improves performance to F1=0.63—surpassing 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 twenty classification task, strong zero-shot models such as `EvaCLIP` reach 62.7% accuracy confirming the dataset's semantic separability and visual grounding. A consistent recall–precision tradeoff across models indicates a generalization bias toward over-retrieval. Overall, SAFIRE's diversity and granularity make it a robust testbed for classification and retrieval-style evaluation under both zero- and low-shot 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 **105K** images across **20** real-world scenarios and **224K** multiple-choice VQA pairs, organized under five groups (natural, industrial, accidental, entertainment, 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, seven open-source MLLMs average **57.5%** MCVQA accuracy, while CLIP-style encoders reach up to **62.7%** in zero-shot classification; adapting the encoder with only **3%** supervision yields a **12.7%** gain. 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.

6 Limitation

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 fixed natural-language prompts (e.g., "A photo of a {scene}"). Exploring prompt diversity, syntactic framing, and semantic variation, along with methods like prompt tuning or renaming (for example, replacing "garbage fire" with "dumpster fire"), could further improve generalization and text–image alignment in safety-critical scenarios.

Second, the multimodal encoders evaluated on SAFIRE, such as CLIP and EvaCLIP, are pretrained on large-scale general-domain image–text corpora. Our domain-specific adaptation was applied only at the encoder level and evaluated on zero-shot and few-shot classification tasks. Domain adaptation for full MLLMs was not explored in this work and remains an important direction for improving multimodal reasoning and robustness in safety-critical fire–smoke scenarios.

Finally, most annotations were generated automatically using a vision–language model and subsequently screened through automated and model-assisted checks. Flagged samples were carefully reviewed by humans to ensure accuracy. This hybrid approach enables large-scale, consistent data creation, though minor linguistic or cultural biases from the source model may still persist. Continued auditing and validation across diverse scenarios, informed by community feedback, will help maintain fairness and reliability.

Overall, these factors represent natural next steps for extending SAFIRE rather than shortcomings of the current benchmark. Addressing them will further enhance its value as a reliable and comprehensive testbed for developing robust multimodal systems in safety-critical environments.

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.

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

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European conference on computer vision (ECCV)*, pages 213–229. Springer.

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.

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*.

Gemma Team. 2025. Gemma 3.

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.

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. 2023. Enhancing real-time fire detection: An effective multi-attention network and a fire benchmark. *Neural Computing and Applications*, pages 1–15.

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, Yeqi Wu, Yujung Cha, Tushar Ajanthan, Gautam Sadeghi, Venkatesh Saligrama, and Yong Jae Lee. 2023a. Visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Hexiang Liu, Yixiao Wang, Peng Zhang, Shaohan Yang, Fanyi Wei, Ming Zhou, Hang Li, Baifeng Chen, Wei-Ying Ma, and Zicheng Liu. 2023b. Eva-clip: Efficient vision-language pre-training with expert knowledge distillation. *arXiv preprint arXiv:2306.17781*.

Meta AI. 2024. Llama-3.2-vision: Vision-enabled instruction-tuned llms. Hugging Face model card. https://huggingface.co/meta-llama/Llama-3.2-11B-Vision.

Mistral AI. 2025. Mistral small 3.1: Versatile 24b parameter multimodal model. Hugging Face model card. https://huggingface.co/mistralai/Mistral-Small-3.1-24B-Instruct-2503.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 779–788.

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.

Xiyang Zhang, Tianheng Fang, Jiemin Wang, Shuai Shao, Zeming Li, Jian Wang, Ping Luo, Gang Sun, and Jifeng Dai. 2022. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*.

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 work are from public sources and contain no personally identifiable or sensitive information. The SAFIRE benchmark was curated to advance research on safety-critical visual reasoning in fire and smoke scenarios and will be released under a CC BY-NC 4.0 license for non-commercial research use. All pretrained models (e.g., Qwen2.5-VL, LLaMA, OpenCLIP) were employed in accordance with their respective open-source licenses. No human subjects or external annotators were involved in data collection or analysis, and therefore no ethical approval was required. Portions of the manuscript were refined using ChatGPT for language clarity; all conceptual and scientific content was solely verified by the authors without AI.

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 "agriculture fire" scenario, we used the combined query terms "agriculture fire", "farm", 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 Dataset Cleaning

Following data collection, we conducted a manual curation phase to refine the dataset. Each candidate scenario was assessed based on two main criteria: (1) overall image quality and noise level, and (2) 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, we applied a **comprehensive manual cleaning process** across all retained images to remove remaining irrelevant or low-quality samples. This reduced the dataset from an initial 118K auto-collected images to **105K high-quality images**, with **12.4%** removed during manual inspection.

Overall, the Multi-Factor Filtering (MFF) strategy significantly improved dataset quality throughout the pipeline by filtering out **37% of irrelevant images** during the early automated stages, discarding **20% of candidate scenarios** due to semantic or visual redundancy, and removing an additional **12.4% of images** during the final manual inspection.

As a result, only **44.2%** 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 224K 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 105K 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., 2025) 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 filtered image is paired with 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.

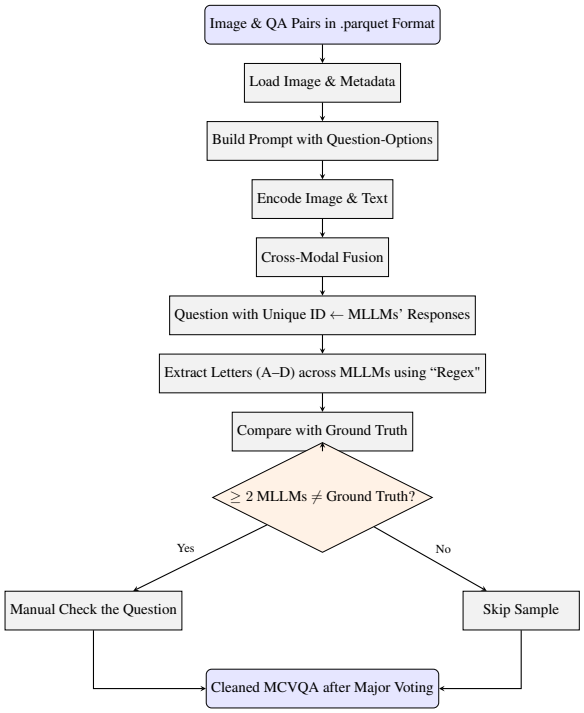

Figure 8: **Majority voting pipeline for cleaning multi-scenario MCVQA answers from multiple MLLMs.**

Output JSON Skeleton

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

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 The Major Voting for Correcting MCVQA

To verify MCVQA quality, we adopt a majority-voting pipeline across multiple MLLMs, visualized in Fig. 8. For each question, model predictions (options A–D) are extracted using regular expressions and compared to the ground-truth answer. If two or more MLLMs disagree with the ground truth, the question is flagged for manual review. Otherwise, the sample is skipped. This strategy efficiently prioritizes human attention toward potentially flawed samples. Notably, the pipeline also identifies questions where model responses differ from available options, indicating unclear labels that need investigation.

D Evaluation Pipeline Algorithms

The full procedural steps for multi-GPU 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 and paired with 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.

In Algorithm 2, the full procedural steps for CLIP-based fire scene classification under both zero-shot and few-shot 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 90%) of the labeled image features to test few-shot capabilities.

E Visual Diversity Across Fire–Smoke Scenarios

Fig. 9 illustrates ensemble examples from SAFIRE's 20 distinct fire and smoke scenarios. While Figures 10, 11, and 12 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 dis-

Algorithm 1: Multi-GPU VQA Evaluation on MCVQA Fire Dataset

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 $p \leftarrow \texttt{build_prompt}(qa)$;
    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 $\hat{y}$;
    Ground Truth $\leftarrow$ 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* $r \in \{1\%, 3\%, 10\%, 50\%, 90\%\}$ **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*

tinctiveness. For example, **Figure 10** presents large-scale natural fires such as *Volcano* and dynamic man-made events like *Rocket Jet Fire* and *Burning Cars*. **Figure 11** features culturally embedded and industrial scenarios including *Torch Fire*, *Steel Forging Fire*, and *Explosion*, each with unique flame structures and textures. **Figure 12** 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.

The scenario order aligns 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, domestic, and behavioral contexts, and help illustrate the generalization challenges faced by both vision-only and multimodal models.

Figure 9: **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) Agricultural Fire, (2) Forest Fire, (3) Volcano, (4) Meteorite/Astronomical Event, (5) Rocket Launch Fire, (6) Gas Flares, (7) Steel Forging, (8) Home Fire, (9) Burning Car, (10) Garbage Trash Fire, (11) Explosion, (12) Firework, (13) Hot Air Balloon, (14) Barbecue, (15) Campfire, (16) Torch Parade, (17) Candlelight Fire, (18) Human Smoking, (19) Incense Burning, (20) Cooking Fire (Gas Stove).

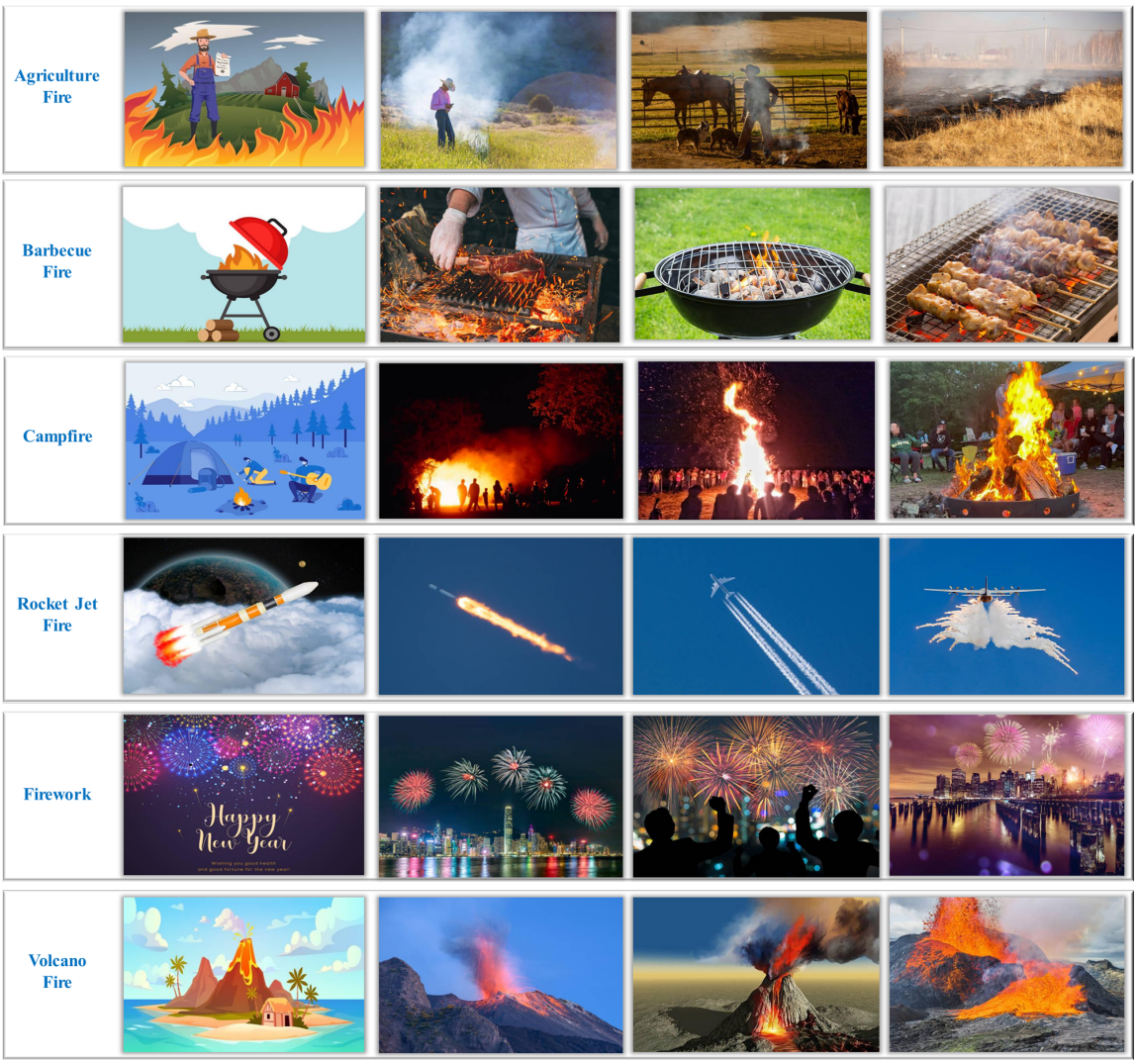

Figure 10: Visual examples of fire-smoke scenarios (number 1-6) used in the benchmark.

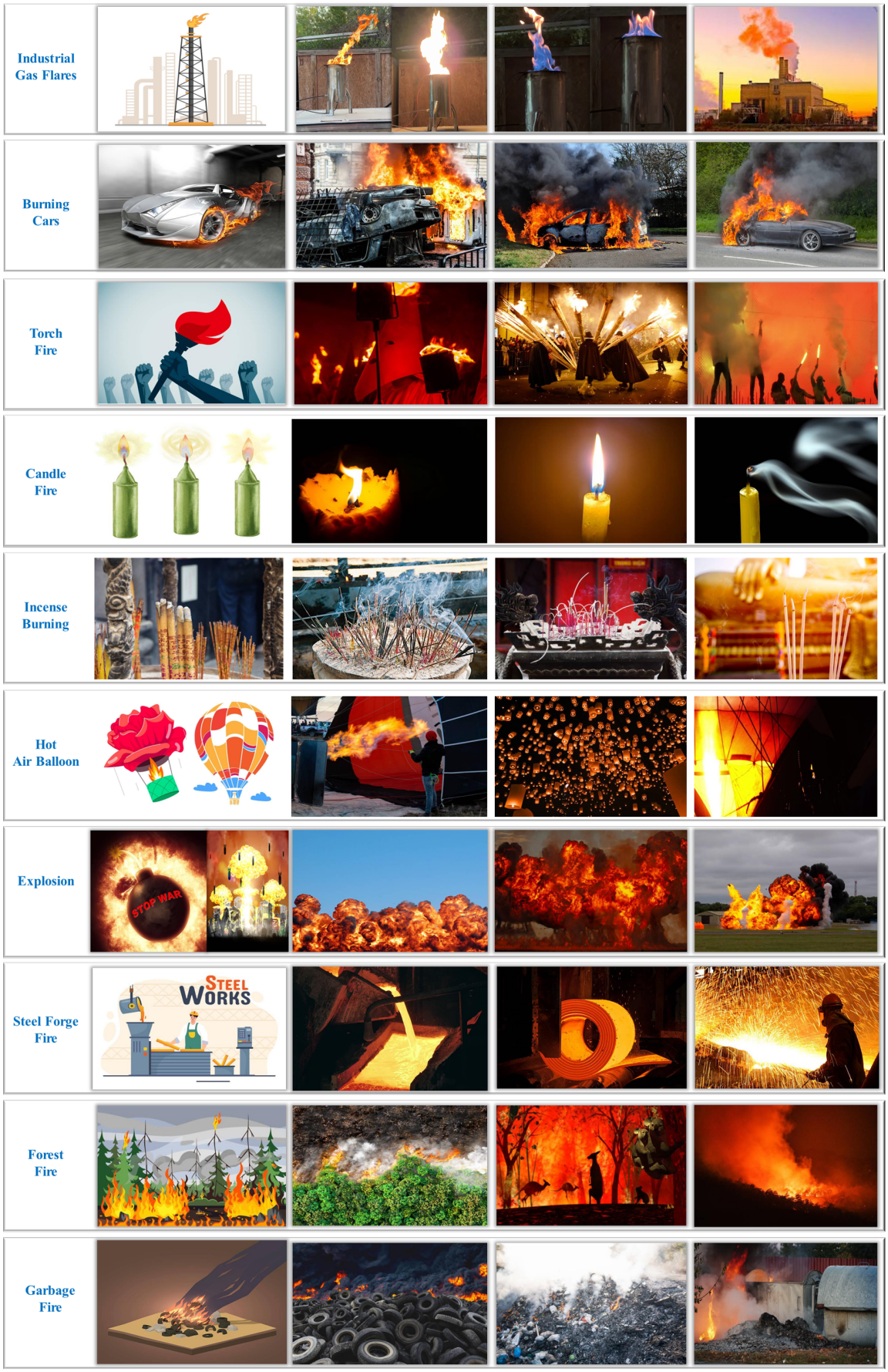

Figure 11: Visual examples of fire-smoke scenarios (number 7-17) used in the benchmark.

Figure 12: Visual examples of fire-smoke scenarios (number 17-20) used in the benchmark.

F A Full Generated MCVQA Example

Example Annotation Overview. The image `volca_10.jpg` in Fig.13 depicts a dramatic volcanic eruption occurring at night.The comprehensive metadata includes both attribute annotations and a set of 20 multiple-choice visual questions that test various levels of multimodal reasoning. These questions assess spatial understanding (e.g., the fire is centered; the smoke is positioned above the flame), 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., the smoke dispersing rather than an explosion). 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 13: Example image from the Volcano scenario.

MCVQA JSON Annotation


```
{
  "Image_name": "volca_10.jpg",
  "caption": "A dramatic volcanic eruption at night, with a bright orange-yellow lava flow
      descending the mountain's slope, accompanied by thick, billowing smoke against a starry
      sky backdrop.",
  "pos": "Center",
  "count": "One",
  "danger level": "Extreme",
  "scenario": "Volcano",
  "fire size": "Massive",
  "smoke presence": "Thick smoke",
  "smoke color": "Gray",
  "flame presence": "Yes",
  "flame color": "Orange/Yellow",
  "object on fire": "Lava flow",
  "negative example detection": "Real fire",
  "possible cause": "Natural volcanic activity",
  "human/animal presence": "No",
  "firefighting presence": "Not applicable",
  "time of day": "Nighttime",
  "Spatial Correlation between Fire and Smoke": "Smoke above the fire",
  "QA_pairs": [
    {
      "question": "What kind of smoke/flame is this?",
      "options": ["A. Industrial fire", "B. Forest fire", "C. Volcano eruption", "D. None of
          the above"],
      "answer": "C. Volcano eruption"
    },
    {
      "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"
    },
```

MCVQA JSON (Continue)


```
    {
      "question": "What object or material is burning?",
      "options": ["A. Building", "B. Vegetation", "C. Vehicle", "D. Not clear"],
      "answer": "D. Not clear"
    },
    {
      "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 fire?",
      "options": ["A. Electrical issue", "B. Vehicle crash", "C. Lightning", "D. Unknown"],
      "answer": "D. Unknown"
    },
    {
      "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": "C. Not applicable"
    },
    {
      "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. Fire expands", "B. Smoke disperses", "C. Fire quenches", "D. Explosion
          occurs"],
      "answer": "B. Smoke disperses"
    },
    {
      "question": "What is the purpose of this fire?",
      "options": ["A. Accident", "B. Celebration", "C. Industrial production", "D. Not sure"],
      "answer": "A. Accident"
    },
    {
      "question": "How would people react to this smoke/fire?",
      "options": ["A. Call for help", "B. Ignore it", "C. Evacuate the area", "D. Not sure"],
      "answer": "C. Evacuate the area"
    },
    {
      "question": "What does 'on fire' mean here?",
      "options": ["A. Literal fire", "B. Excited situation", "C. Under attack", "D. Not
          Applicable"],
      "answer": "A. Literal fire"
    }
  ]
}
```