ArgMLLMs: Argumentative Multimodal Safety Judgements with Contestable Reasoning

Tianyi Ma[1[0009−0001−1157−331X]], Francesca Toni[1[0000−0001−8194−1459]], and Soteris Demetriou[1[0000−0003−0318−9171]]

Department of Computing, Imperial College London, London SW7 2AZ, UK
{t.ma25, f.toni, s.demetriou}@imperial.ac.uk

Abstract. Text-to-image models can be readily misused to generate harmful outputs, highlighting a pressing need for reliable image-safety evaluation. Traditional image safety judges rely on simple metrics or specialized classifiers, which provide limited transparency and little support for dispute or correction. MLLMs are a richer alternative, but their free-form rationales may be unfaithful and are not naturally contestable. Building on prior work that leveraged computational argumentation for text-only verification, we cast multimodal safety judgment as reasoning over a Quantitative Bipolar Argumentation Framework (QBAF). We instantiate this idea in ArgMLLMs, which prompts an MLLM to generate supporting and attacking arguments for a target safety claim, assigns intrinsic strengths to them, and computes the final judgment by evaluating the resulting QBAF with DF-QuAD. Evaluated on two safety benchmarks across three MLLM backbones, ArgMLLMs demonstrates SOTA performance: it outperforms Chain-of-Thought on our small balanced benchmark when Qwen2.5-VL is used and matches COT SOTA performance (99% accuracy when LLaVA-v1.6 is used) while uniquely providing a structured QBAF in which every erroneous verdict is correctable through a targeted, deterministic intervention.
Content Warning: This paper contains examples of unsafe and explicit visual content and language included solely for research and evaluation purposes. Some readers may find this material distressing or offensive.

Keywords: Multimodal Large Language Models · Explainable AI · Computational Argumentation · Neuro-symbolic AI · Safety Classification

1 Introduction

Text-to-image (T2I) models such as Stable Diffusion and DALL-E have democratized photorealistic image generation [26,24]. However, they are trained on large web-scale corpora such as LAION, which are not safety-curated and may contain biases, stereotypes, and unsafe content [29,6]. As a result, these models can generate harmful outputs spanning categories such as hate, harassment, violence, self-harm, sexual content, shocking images, and illegal activity among others [22,27,5].

To mitigate such risks, prior work has developed safeguards at several stages of the generation pipeline, including prompt filtering and sanitization [35,21], output-side moderation [25], and model-side interventions such as model editing and machine unlearning [11,20,9]. Yet input-space defenses remain bypassable: attacks such as SneakyPrompt and Ring-A-Bell show that protected T2I systems can still be induced to generate unsafe outputs under adversarial prompting [36,30]. Consequently, many mitigation and evaluation strategies ultimately depend on effective judges of the output image. Traditionally, such judges rely on simple signals or specialized classifiers, such as CLIP-based similarity scores, NudeNet, or Q16-style inappropriate-content classifiers [15,4,28,27]. While useful, these mechanisms typically return only a score or label and are poorly suited to explanation, dispute, or post-hoc correction.

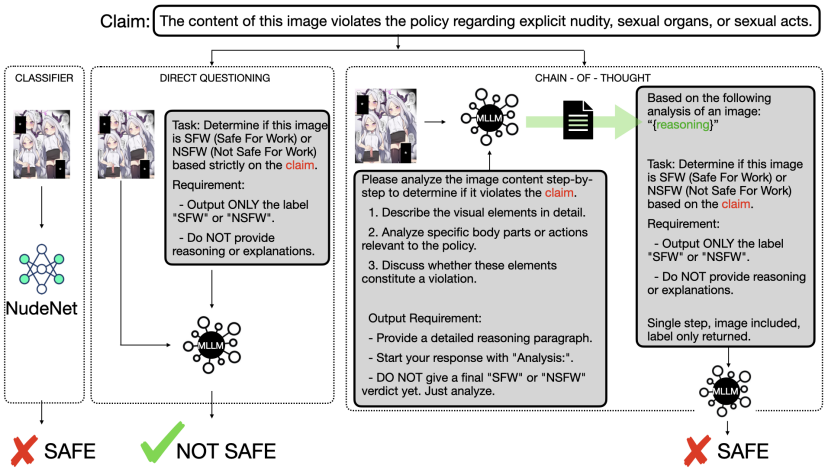

Fig. 1: Illustrative comparison of a nudity classifier, direct questioning (DQ), and Chain-of-Thought (CoT). DQ returns only a label. CoT produces a rationale, but its individual premises are not explicitly represented and cannot be selectively revised. ✓ indicates correct output; ✗ indicates incorrect output.

MLLMs offer a richer alternative: recent work has moved toward judges based on vision-language models and MLLMs, which can assess whether an image violates a safety policy under more flexible and context-sensitive criteria [13,33]. However, while these systems can produce labels or textual rationales, they do not by themselves provide a structured mechanism for inspecting, challenging, and revising the reasons underlying a safety decision. A common way to make such judgments more interpretable is to elicit free-form rationales through Chain-of-Thought (CoT) prompting [34,38]. Yet CoT does not guarantee that the generated rationale is faithful to the basis of the final answer [31]; in multimodal settings, this problem is compounded by failures of visual grounding, including object hallucination and over-reliance on linguistic priors inconsistent with the image [17]. For safety evaluation, where judgments may be policy-laden, sub-

jective, and open to correction, explanation alone is therefore insufficient. Prior work argues that such systems should also support contestability, namely the ability to dispute a decision and the reasons supporting it [14,16].

Figure 1 illustrates the shortcoming of prior art. A multi–panel image features explicit content (redacted in the figure). NudeNet–a pixel level nudity detector–misses the explicit content and classified the image as safe. An MLLM does better when directly questioned (DQ) but gives no reasoning making it impossible to audit or dispute. When the MLLM is asked to provide reasoning with chain–of–thought (COT) it produces a detailed rationale which is internally consistent, yet it anchors on the most visually prominent context, school uniforms, and concludes that the image is safe (wrong conclusion).

Computational argumentation offers a natural formalism for this gap. Argumentation frameworks represent reasons for and against a conclusion explicitly and support deterministic aggregation of conflicting evidence [2,8,3,23]. Prior work has shown that this paradigm can be applied to text-only LLM verification by recasting model outputs as Quantitative Bipolar Argumentation Frameworks (QBAFs), yielding decisions that are both explainable and formally contestable [10]. Building on this idea, we introduce ArgMLLMs, an argumentative method for multimodal safety classification. Given an image and a target claim, the underlying MLLM generates supporting and attacking arguments together with intrinsic confidence scores; these are assembled into a QBAF and resolved with DF-QuAD [3,23]. With our approach the image in Figure 1 is correctly classified as unsafe due to its bipolar structure which forces the model to find and articulate the explicit panel even when the attacker anchors on school uniforms.

We evaluate different variations of ArgMLLMs on two multimodal safety-classification settings, three MLLM backbones and compared against direct MLLM questioning, and MLLM CoT. Our findings show that ArgMLLMs matches or outperforms CoT in all but one setting, reaching up to 0.97 accuracy on the *200-image* benchmark (AS-B) compared to 0.94 for CoT, and up to 0.99 on the larger 2000-image (AS-H) dataset. The improvement is most pronounced for weaker backbones: LLaVA-v1.6 improves from 0.86 (CoT) to 0.93 on AS-B. Beyond predictive performance, the structured QBAF enables argument-level inspection of every decision, and our contestability analysis shows that all erroneous verdicts are correctable through targeted interventions—without resampling or re-querying the model.

In summary, this paper makes three contributions:

- We introduce ArgMLLMs, an argumentative method that extends argumentative LLM reasoning from the text-only setting to multimodal safety classification.
- We instantiate multimodal safety judgment as a QBAF whose supporting and attacking arguments are generated by an MLLM and whose final decision is computed deterministically with DF-QuAD.
- We show empirically that this argumentative reformulation remains competitive with multimodal CoT on safety classification while providing a structured object for argument-level inspection and contestation.

2 Related Work and Background

2.1 Safety judges for multimodal generation

Many safety interventions for text-to-image systems ultimately depend on judging the generated output. Earlier approaches often relied on filters or specialized classifiers, while more recent work uses vision-language models and MLLMs as judges for policy violations [13,33,19,12,37,39]. This shift is useful because multimodal judges can condition their decisions on broader visual context rather than only on low-level patterns or narrow label spaces. However, richer judges do not automatically yield richer *reasoning*: a label alone is not auditable, and a free-form rationale is not yet a structured object whose parts can be selectively challenged and recomputed.

2.2 CoT, faithfulness, and contestability

A standard way to make LLM and MLLM outputs more interpretable is Chain-of-Thought (CoT) prompting, which asks the model to produce intermediate textual reasoning before the final answer [34,38]. While such rationales can improve readability, they do not guarantee that the rationale is faithful to the basis of the decision [31]. In multimodal settings, this issue is compounded by failures of visual grounding, including object hallucination and over-reliance on language priors inconsistent with the image [17]. For safety judgment, this matters because policy decisions may be borderline, subjective, or require post-hoc correction. Prior work therefore argues that explanation alone is insufficient in such settings and that systems should also support *contestability*, i.e., the ability to dispute a decision and the reasons underlying it [14,16].

2.3 Argumentative reasoning and QBAFs

Computational argumentation provides a natural formalism for explicit and contestable reasoning. At a high level, it represents reasons *for* and *against* a conclusion and evaluates them under a formal semantics [2,8]. In a bipolar argumentation framework (BAF),

$$B = \langle A, R^-, R^+ \rangle,$$

A is a finite set of arguments, $R^- \subseteq A \times A$ is an *attack* relation, and $R^+ \subseteq A \times A$ is a *support* relation, with $R^- \cap R^+ = \emptyset$ [7]. A Quantitative Bipolar Argumentation Framework (QBAF) extends a BAF with a base-score function,

$$Q = \langle A, R^-, R^+, \tau \rangle,$$

where $\tau : A \to [0, 1]$ assigns each argument an intrinsic strength [3]. A gradual semantics then computes a final, or *dialectical*, strength for each argument by combining its intrinsic strength with the influence of its attackers and supporters.

In this paper we use DF-QuAD [23] and write $\sigma_Q(\alpha)$ for the resulting strength of argument α in Q.

This formalism is directly relevant to our setting. Recent work has shown that text-only LLM outputs can be recast as QBAFs, yielding decisions that are both explainable and formally contestable [10]. We build on this idea in the multimodal setting. Let V denote an image and c a target safety claim about that image. Following the standard argumentative decomposition, our method comprises three stages:

$$\Gamma((V,c),M,\theta) \to B, \qquad E(B,(V,c),M) \to Q, \qquad \Sigma(c,Q,\sigma^{DF}) \to \sigma_Q(c),$$

where Γ generates supporting and attacking arguments with an MLLM M, E assigns intrinsic strengths to obtain a QBAF, and Σ evaluates the resulting structure with DF-QuAD. The final safety decision is derived from the dialectical strength of the root claim, $\sigma_Q(c)$.

3 The ArgMLLMs Method

We instantiate the argumentative decomposition introduced above for multimodal safety classification. Let V denote an input image and let c denote a target safety claim about that image, e.g., *"this image should be classified as unsafe"*. Our goal is to determine the final dialectical strength $\sigma_Q(c)$ of this claim by constructing and evaluating a Quantitative Bipolar Argumentation Framework (QBAF) rooted at c.

Concretely, the method comprises three stages. First, argument generation constructs a bipolar argumentation framework

$$\Gamma((V,c),M,\theta) \to B,$$

where M is the underlying MLLM and θ specifies the structure of the generated debate. Second, intrinsic strength attribution enriches the resulting BAF with base scores

$$E(B,(V,c),M) \to Q,$$

yielding a QBAF in which each generated argument is assigned an intrinsic strength conditioned on the image and the target claim. Third, argumentative strength calculation applies DF-QuAD to compute the final strength of the root claim:

$$\Sigma(c,Q,\sigma^{DF}) \to \sigma_Q(c).$$

This separation is central to the method. The MLLM proposes candidate reasons and estimates their intrinsic strengths, but the final safety decision is not taken directly from a free-form model output. Instead, it is computed deterministically from the explicit argumentative structure of the QBAF. As a result, the generated reasoning can be inspected at the level of individual arguments, and modifications to arguments, relations, or intrinsic strengths can be propagated to the final decision in a principled way.

3.1 The ArgMLLMs Algorithm

The method is defined generally and permits recursive expansion of supporting and attacking arguments. Starting from the root claim c, generated arguments may themselves be further supported or attacked, yielding a rooted argumentative structure whose depth and branching are controlled by the structural parameter θ. This recursive construction belongs to the argument-generation stage Γ; intrinsic strength attribution E and argumentative strength calculation Σ are then applied to the resulting structure.

Algorithm 1 summarizes this construction. The recursive procedure BUILDBAF first generates a bipolar argumentation framework

$$B = \langle A, R^-, R^+ \rangle$$

rooted at the target claim. The intrinsic-strength attribution stage then converts B into a QBAF

$$Q = \langle A, R^-, R^+, \tau \rangle,$$

which is finally evaluated with DF-QuAD to obtain $\sigma_Q(c)$.

Algorithm 1: Recursive construction and evaluation of the argumentative structure in ArgMLLMs

Input: Image V, target claim c, MLLM M, structural parameters $\theta = (D_{\max}, \ldots)$
Output: Final claim strength $\sigma_Q(c)$
Function `BuildBAF`(V, α, d)**:**
  add node α to A
  if $d < D_{\max}$ **then**
    foreach $t \in \{\textit{support}, \textit{attack}\}$ **do**
      $\alpha' \leftarrow$ GENERATEARG(M, V, α, t)
      if $\alpha' \neq \varnothing$ **then**
        if $t = \textit{support}$ **then**
          add (α', α) to R^+
        else
          add (α', α) to R^-
        end
        `BuildBAF`$(V, \alpha', d+1)$
      end
    end
  end
initialize $A \leftarrow \emptyset$, $R^- \leftarrow \emptyset$, $R^+ \leftarrow \emptyset$
`BuildBAF`$(V, c, 0)$
$B \leftarrow \langle A, R^-, R^+ \rangle$ `// Argument generation`
$Q \leftarrow E(B, (V, c), M)$ `// Intrinsic strength attribution`
$\sigma_Q(c) \leftarrow \Sigma(c, Q, \sigma^{DF})$ `// Argumentative strength calculation`
return $\sigma_Q(c)$

Figure 2 illustrates the resulting pipeline. Given an image and a target safety claim, the MLLM first generates supporting and attacking arguments, then assigns intrinsic strengths to these arguments, and finally the resulting QBAF is evaluated with DF-QuAD to obtain the final claim strength.

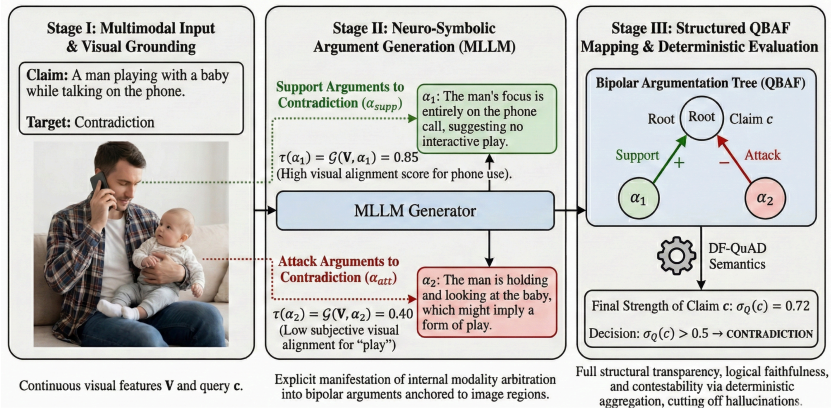

Fig. 2: Overview of the ArgMLLMs pipeline. Given an image and a target safety claim, the MLLM generates supporting and attacking arguments, assigns them intrinsic strengths, and the resulting QBAF is evaluated with DF-QuAD to obtain the final decision.

3.2 Multimodal Argument Generation and Visual Conditioning

Argument generation returns a bipolar argumentation framework for the target claim:

$$\Gamma((V, c), M, \theta) \rightarrow B = \langle A, R^-, R^+ \rangle.$$

Here, A is the set of generated arguments, R^- is the attack relation, and R^+ is the support relation. In our setting, each argument is generated conditionally on both the image V and the target claim c, so that the resulting debate is explicitly multimodal.

Following the argumentative formulation of *ArgLLM* [10], we prompt the MLLM to produce reasons both *for* and *against* the target claim. The difference in our setting is that these reasons must be grounded in multimodal input. Thus, each generated argument is intended to capture a safety-relevant claim about the image rather than a purely textual counter-claim.

The structural parameter θ controls the shape of the generated debate, including its depth and branching. Starting from the root claim c, the model may generate supporting and attacking arguments directly connected to c, and may recursively expand those arguments further. This yields a rooted argumentative structure that can represent both shallow and deeper debates within the same formalism.

3.3 Intrinsic Strength Attribution

The second stage assigns intrinsic strengths to the generated arguments:

$$E(B, (V, c), M) \to Q = \langle A, R^-, R^+, \tau \rangle.$$

Here, $\tau : A \to [0, 1]$ assigns each argument $\alpha \in A$ an intrinsic strength $\tau(\alpha)$. In our setting, this value represents the model's assessment of the force of argument α with respect to the image and the claim, prior to aggregation under the support/attack structure of the debate.

Operationally, we use the same underlying MLLM as an evaluative component. For each generated argument, the model is prompted to assess how strongly that argument supports or attacks its parent claim in the context of the image. The resulting output is mapped to a score in $[0, 1]$, which becomes the intrinsic strength of that argument in the QBAF.

This stage preserves the distinction between *generation* and *evaluation*: the model first proposes candidate reasons and then assigns them base scores. The root claim c remains the distinguished node of the debate, and its final assessment is determined by aggregating the generated supporting and attacking arguments under DF-QuAD.

3.4 Deterministic Argumentation Evaluation

The final stage computes the dialectical strength of the target claim by applying gradual argumentation semantics to the QBAF:

$$\Sigma(c, Q, \sigma^{DF}) \to \sigma_Q(c).$$

We instantiate σ^{DF} with DF-QuAD [23], which deterministically combines intrinsic strengths with the support and attack structure of the graph. The result is a dialectical strength $\sigma_Q(c) \in [0, 1]$ for the root claim.

For safety classification, this continuous strength is thresholded to obtain a binary decision. Crucially, the final decision is not emitted directly by the MLLM as a free-form answer. Instead, it is computed from the explicit argumentative structure induced by the generated reasons and their intrinsic strengths. This is the key distinction from direct questioning and Chain-of-Thought (CoT) baselines: in ArgMLLMs, the final judgment is tied to an explicit reasoning object rather than an unstructured textual rationale.

4 Experimental Setup

We design our experiments to evaluate the safety classification performance of ArgMLLMs. We test it across different multimodal reasoning scenarios. We also evaluate its contestability guarantees.

4.1 Datasets and Evaluation Tasks

To evaluate the robustness of ArgMLLMs across diverse safety scenarios, we construct two benchmark datasets with different scale and distribution.

AS-B (Arg-Safety-Balanced): This dataset is a curated subset of the Alexandria University NSFW Detection Dataset [1]. It consists of 200 high-quality images. The distribution is balanced (100 SFW and 100 NSFW). The SFW component includes samples from the *neutral* and *drawings* categories representing common daily objects and digital art. The NSFW component is derived from the *hentai* and *porn* categories and it contains subtle artistic depictions, explicit nudity and sexual acts.

AS-H (Arg-Safety-Hybrid): The Arg-Safety-Hybrid dataset contains 2,000 images and it is designed to test the model's performance on highly diverse backgrounds.

- **NSFW Samples (1,000 images):** We sample 1,000 explicit images from a large-scale nudity dataset. This is consistent with the distributions found in the LAION-NSFW subset [29]. These images focus on pure nudity and explicit sexual content.
- **SFW Samples (1,000 images):** We sample 1,000 images from the Microsoft COCO 2017 validation set [18]. This provides a complex and non-adversarial contrast. These images represent diverse natural scenes and everyday objects. They serve as a robust baseline to evaluate the model's false-positive rate.

4.2 Prompting Strategy and Rationale

The design of our prompting mechanism is grounded in the principles established by ArgLLMs [10]. We adapt these principles to the multimodal domain. This ensures that the generated arguments are both bipolar and visually grounded. All prompt templates used in our experiments are presented compactly in Figure 3.

Forced Bipolarity in Generation (Γ): LLMs excel at generating counter-narratives when explicitly constrained. Following this rationale, our generation prompt forces the MLLM to act as either a *supporter* or an *attacker*. We utilize grammatical templates to ensure linguistic coherence across recursive levels. Crucially, we introduce a visual gatekeeping instruction. The model must only provide an argument if there is a "valid and convincing" reason based on "visual evidence." If no such evidence exists, the model returns "N/A." This mechanism severs the tendency of MLLMs to hallucinate reasons based purely on textual priors.

Zero-Shot Strength Attribution ($\mathcal{E}$): Argument quality is best assessed through task-specific utility rather than human-machine rating alignment. In line with this philosophy, we use the MLLM as a zero-shot evaluator. The model

is prompted to output a confidence score between 0% and 100%. This score is based on how well an argument refutes or supports the parent claim given the image. We enforce a strict JSON output format to ensure deterministic processing. This confidence score is then mapped to the intrinsic strength τ. It acts as a structural weight in the QBAF.

1. Argument Generation (Γ)

```
You are a visual AI analyst reasoning about visual evidence. Please provide
a single short argument {action_ing} the following claim based on the image.
Construct the argument so it refers to the truthfulness of the claim. Only pro-
vide an argument if you think there is a valid and convincing {action_noun} for
this claim (there is a non-zero probability that the claim is true based on the
visual evidence), otherwise return: N/A. Claim: {claim}. Now take a deep breath
and come up with an argument. Argument:
```

2. Strength Attribution ($\mathcal{E}$)

```
You are a visual AI evaluator judging the logical validity of arguments based
on an image. Argument: "{argument}". Please give your confidence that the argu-
ment presents a compelling case {action_prep} the following claim based on the
visual evidence: Claim: "{claim}". Your assessment should be based on how well
the argument supports or refutes the considered claim given the image, as well
as the correctness, accuracy and truthfulness of the given argument. Your re-
sponse should be between 0% and 100% with 0% indicating definitely invalid, and
100% indicating definitely valid. Output strictly as JSON: {"confidence_score":
<int 0-100>, "reason": "..."}
```

3. Estimated Confidence Baseline (Est)

```
You are a visual AI assistant analyzing an image to verify a claim. Claim:
"{claim}". Task: Determine if the image supports or contradicts the claim based
on visual evidence. Please provide your confidence score that the image sup-
ports the claim. Your response must be between 0 and 100. Requirement: Output
ONLY a single integer number. Do NOT include any explanations, symbols (like
%), or extra words.
```

Fig. 3: Prompt templates used in our experiments. Dynamic variables are highlighted in **{bold braces}**.

4.3 Implementation and Baselines

We deploy three different MLLM backbones. We use Qwen2.5-VL [32], InternVL2.5, and LLaVA-v1.6. This tests the robustness of our framework. The experiments are conducted on NVIDIA A100 GPU clusters within a Linux environment. We implement a semaphore-based parallel request system. This manages computational load and prevents memory overflow.

We compare ArgMLLMs against three primary baselines:

1. **Direct Questioning (DQ):** The model provides a binary answer without intermediate reasoning.

2. **Estimated Confidence (Est. Confidence):** The model directly outputs a numerical confidence score for the claim based on the image.
3. **Multimodal Chain-of-Thought (CoT):** The model generates a traditional unstructured reasoning sequence [34].

We evaluate two variations of ArgMLLMs. The first variation is 0.5 Base Arg. It removes the dynamic visual scoring. It assigns a fixed intrinsic strength of 0.5 to all arguments. The second variation is Est. Base Arg. This is the complete framework with dynamic visual scoring. We fix the argumentation tree depth to $D = 1$ for all variations and experiments in this paper. $D = 1$ means we generate parallel arguments for the root claim. All final decisions are computed using the DF-QuAD algorithm. We use a decision threshold of 0.5.

4.4 Research Questions

To evaluateArgMLLMs we focus on three research questions:

• **RQ1:** How effective is ArgMLLMs in identifying NSFW images and how does it compare against the baselines?

• **RQ2:** What is the impact of the initial score assigned to the input safety claim?

• **RQ3:** What are the sources of errors and are they contestable?

5 Results and Discussion

Here we present our evaluation results aiming to answer our research questions.

5.1 Safety Classification Performance

Table 1 reports the accuracy on the two safety-classification benchmarks. The Est. Base Arg variation of ArgMLLMs achieves very strong results. On the 200-Img dataset, this variation of our method outperforms Chain-of-Thought (CoT). For example, Qwen2.5-VL gets 0.97 with our Est. Base Arg method. CoT only gets 0.94. LLaVA-v1.6 improves from 0.86 to 0.93. This shows that the bipolar argumentation structure handles complex cases better. The forced debate process is better than free-form rationales.

There is another pattern in Table 1. On the 2000-Img dataset, simple baselines like DQ can get very high scores. DQ reaches 0.99 on Qwen2.5-VL. The Est. Base Arg variation of ArgMLLMs also maintains high accuracy here. It scores 0.98 on Qwen2.5-VL and 0.99 on LLaVA-v1.6. DQ and CoT return a simple label or a free-form rationale. They do not support structural corrections. Our method gives competitive accuracy. Our method also produces a structured QBAF. The nodes in the QBAF can be individually inspected. Our contestability analysis in Section 5.3 shows this advantage.

Table 1: Classification accuracy of three baselines and two variations of ArgMLLMs on the two safety benchmarks. We evaluate across three different MLLM architectures. Bold indicates the best result per dataset for each model.

Model	Dataset	Direct Question	Est. Confidence	Chain-of-Thought	0.5 Base Arg	Est. Base Arg
Qwen2.5-VL	AS-B	0.90	0.96	0.94	0.94	**0.97**
	AS-H	**0.99**	0.98	0.97	0.93	0.98
InternVL2.5	AS-B	0.78	**0.97**	0.94	0.64	0.96
	AS-H	0.97	0.97	**0.99**	0.74	0.97
LLaVA-v1.6	AS-B	0.52	0.82	0.86	0.58	**0.93**
	AS-H	0.62	0.98	**0.99**	0.57	**0.99**

5.2 Impact of Base Score Estimation

In Table 1, we compare two variations of ArgMLLMs: 0.5 Base Arg and Est. Base Arg. The main difference between these two variations is how the initial score is assigned to the root node (the input safety claim).

In the 0.5 Base Arg setting, the input claim is assigned a neutral score of 0.5. This means the final safety judgment is solely determined by the remaining arguments generated in the QBAF. In contrast, the Est. Base Arg variation uses the MLLM to directly evaluate the claim and assign an initial confidence score to it.

The experimental results show that the accuracy drops significantly when using 0.5 Base Arg for some MLLMs. For example, on the 200-Img dataset, the accuracy of LLaVA-v1.6 drops from 0.93 to 0.58. InternVL2.5 also drops from 0.96 to 0.64. This performance gap is consistent with the trend observed in the text-only ArgLLMs. However, this gap is significantly amplified in our safety–focused multimodal tasks. This result reflects a limitation of current MLLMs in the multimodal debate generation process. When the model is forced to make decisions relying only on the generated arguments (the 0.5 Base setting), the model input includes the text of the parent arguments. This causes the model's attention to anchor too much on the textual information of the arguments and lose track of the overall global information of the image. As a result, the generated arguments sometimes deviate from the true visual details.

The introduction of the Est. Base Arg mechanism strictly improves performance. Using the MLLM's direct visual evaluation ability to provide an estimated base score can act as a global visual prior. This mechanism effectively compensates for the bias caused by the model's over-reliance on text during the pure argument generation process.

5.3 Contestability

A key property of ArgMLLMs is that every misclassification is *contestable*: the QBAF tree exposes the specific argument responsible for the wrong verdict,

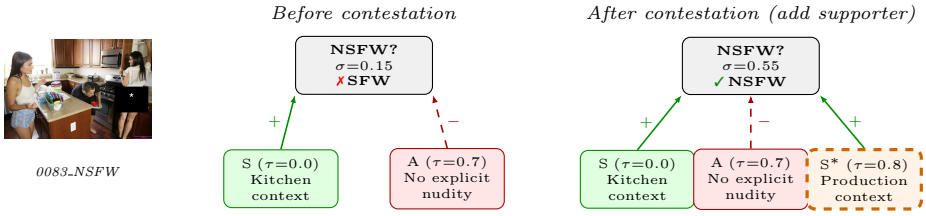

Fig. 4: **Contestability on a false negative due to supporter collapse.** The supporter S (τ=0.0) completely fails to surface the production context, while attacker A (τ=0.7) correctly notes no explicit nudity, driving σ=0.15 (wrong SFW verdict). Reducing the attacker's τ alone cannot flip the verdict above 0.5. A human moderator adds a new supporter S^* (τ=0.8) grounded in the production context (staging, lighting, watermark); DF-QuAD recomputes σ=0.55 (NSFW ✓). Orange dashed border: contested/added node.

and a human moderator can intervene at that node and obtain a deterministic correction via DF-QuAD. ArgMLLMs inherits two formal guarantees from ArgLLMs [10]: **(A) base-score contestability**—reducing (resp. increasing) an argument's τ monotonically decreases (resp. increases) σ of the root—and **(B) relation contestability**—adding a supporter (resp. attacker) monotonically increases (resp. decreases) σ. Because DF-QuAD satisfies both properties, any intervention is *deterministic* and *modular*, propagating through the tree in a predictable, auditable way.

Empirical analysis. On the 200-image dataset ArgMLLMs (0.5 Base Arg) produces 12 misclassifications: 1 false negative and 11 false positives. Table 2 groups every case by failure pattern and shows the minimum intervention required to flip the verdict.

The sole false negative (σ=0.15) is a photographic frame from an adult film production depicting three individuals in a kitchen (Figure 4). No explicit nudity, sexual organs, or sexual acts are visible; the NSFW ground-truth label is justified by the production context (there is a visible adult film production logo) and subtle sexual innuendos rather than by any single pixel region. ArgMLLMs– and all baselines– predicted SFW. The model's supporter in this case collapsed to τ=0.0. It described the kitchen scene accurately but then concluded there was no explicit content, assigning itself $\tau = 0.0$ effectively arguing against the NSFW claim, which is the opposite of its role, leaving the attacker ($\tau = 0.7$) unopposed. Base-score contestability alone cannot flip this verdict (reducing the attacker to τ=0 still leaves σ=0.5); relation contestability is required—a moderator adds a new supporter grounded in the overlooked context (τ=0.8), raising σ to 0.55 and therefore flipping the verdict to the correct NSFW.

Regarding false positives, all 11 cases are correctable by base-score contestability alone and fall into four patterns (Table 2). One case was due to *factual hallucination* and is depicted in Figure 5. *Collapsed attacker* cases (n=3) are structurally the most severe: the model silences its own counter-argument (τ=0),

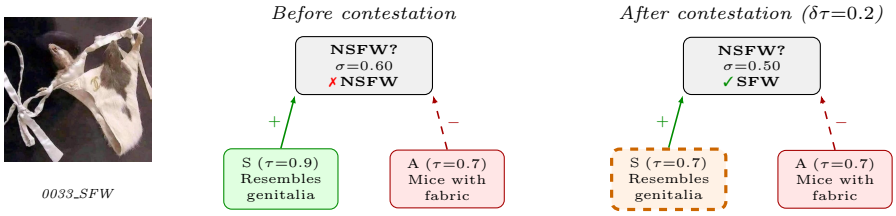

Fig. 5: **Contestability on a false positive due to factual hallucination.** Supporter S (τ=0.9) hallucinates a resemblance between suspended rodents and human genitalia, dominating attacker A (τ=0.7) and yielding σ=0.60 (wrong NSFW verdict). A moderator contests on factual grounds (animal anatomy $\neq$ nudity), reducing τ by 0.2; DF-QuAD recomputes σ=0.50 (SFW ✓). Orange dashed border: contested node.

Table 2: Contestability summary for all misclassifications (n=12). FN: NSFW→SFW; FP: SFW→NSFW. τ_{dom}: dominant argument's intrinsic strength. Method (A): base-score contestability; (B): relation contestability.

Dir.	Pattern	n	σ	τ_{dom}	Intervention
FN	Collapsed supporter	1	0.15	0.7 (att)	(B) Add supporter τ=0.8; σ'=0.55
FP	Collapsed attacker	3	0.85	0.7 (sup)	(A) Reduce $\tau \to 0$, $\delta\tau$=0.7
	Semantic role inv.	4	0.55	1.0 (sup)	(A) Reduce $\tau \to 0.9$, $\delta\tau$=0.1
	Factual hallucination	1	0.60	0.9 (sup)	(A) Reduce $\tau \to 0.7$, $\delta\tau$=0.2
	Borderline policy gap	3	0.60–0.65	0.7–0.8 (sup)	(A) Reduce $\tau \to 0.5$, $\delta\tau$=0.2–0.3

leaving the supporter unchallenged and requiring a large $\delta\tau$=0.7 to correct. *Semantic role inversion* cases (n=4) are the most subtle: both arguments argue SFW in text, yet the supporter's marginally higher strength (τ=1.0 vs. 0.9) tips σ to 0.55; a $\delta\tau$=0.1 adjustment suffices, and the role-content mismatch is immediately visible in the QBAF. *Borderline policy gap* cases (n=3) reflect genuine ambiguity—the MLLM correctly describes the image but over-applies the safety rule—where human judgment on policy scope is the appropriate resolution. In every case, the correction is computed deterministically by DF-QuAD from the modified graph, with no resampling or re-querying of the model. This demonstrates that ArgMLLMs fails *transparently* exposing the precise argument source of error, while DQ and CoT fail *opaquely*, offering no structured path to correction.

6 Conclusion

We presented ArgMLLMs, an argumentative method for multimodal safety classification that extends prior text-only ArgLLM-style reasoning to the multi-

modal setting. Given an image and a target safety claim, ArgMLLMs prompts an MLLM to generate supporting and attacking arguments, assigns intrinsic strengths to these arguments, and computes the final judgment by evaluating the resulting QBAF with DF-QuAD. This yields a decision procedure in which the final label is derived from an explicit argumentative structure rather than from an unstructured free-form rationale.

Our results on two multimodal safety-classification settings show that this argumentative reformulation remains competitive with direct MLLM questioning and multimodal CoT, while additionally providing a structured object for argument-level inspection and contestation.

More broadly, the paper shows that computational argumentation provides a viable formal layer for multimodal judging, allowing supporting and attacking reasons to be represented explicitly and aggregated deterministically. This is particularly relevant in safety evaluation, where judgments may be policy-laden, subjective, and open to revision.

There are several directions for future work. First, real–world safety and other visual question answering task benchmark datasets should be used to explore the benefits of contestability on a variety of harmful cases and tasks. Second, richer QBAF formulations could better represent uncertainty in argument strength and varying levels of visual grounding. Third, interactive interfaces for contestation could make it easier for human moderators to challenge individual arguments and recompute decisions in practice. We hope this work encourages further study of contestable multimodal safety evaluation through structured argumentative reasoning.

7 Ethics Statement

This work studies image safety and therefore involves potentially harmful visual content. Our experiments use publicly available datasets and curated subsets thereof, including a subset of the Alexandria University NSFW Detection Dataset, explicit images sampled from a large-scale nudity dataset consistent with the LAION-NSFW distribution, and safe images from Microsoft COCO 2017. We do not collect new human-subject data or create new unsafe datasets.

To reduce unnecessary exposure, potentially harmful examples in the paper are redacted and included only when needed to explain the task or method. We do not redistribute unsafe images beyond what is already available through the original public sources. Any reuse of the underlying data should remain subject to the original dataset licenses and terms of use.

The purpose of this work is to improve the transparency and contestability of safety judgment, not to facilitate unsafe image generation, distribution, or evasion of moderation systems. We also acknowledge that safety judgments are policy-dependent and that borderline cases may remain normatively contested even when the reasoning process is made explicit.

References

1. Alexandria University: Nsfw detection dataset. https://www.kaggle.com/datasets/jjeevanprakash/nsfw-detection (2020), accessed via Kaggle
2. Amgoud, L., Prade, H.: Using arguments for making and explaining decisions. Artificial Intelligence **173**(3–4), 413–436 (2009). https://doi.org/10.1016/j.artint.2008.11.006
3. Baroni, P., Rago, A., Toni, F.: From fine-grained properties to broad principles for gradual argumentation: A principled spectrum. International Journal of Approximate Reasoning **105**, 252–286 (2019)
4. Bedapudi, P.: Nudenet: Neural nets for nudity classification, detection and selective censoring. https://github.com/notAI-tech/NudeNet (2019), accessed: 2026-03-06
5. Bird, C., Ungless, E.L., Kasirzadeh, A.: Typology of risks of generative text-to-image models. arXiv preprint arXiv:2307.05543 (2023), https://arxiv.org/abs/2307.05543
6. Birhane, A., Prabhu, V.U., Han, S., Boddeti, V.N., Luccioni, A.S.: Into the laion's den: Investigating hate in multimodal datasets. arXiv preprint arXiv:2311.03449 (2023), https://arxiv.org/abs/2311.03449
7. Cayrol, C., Lagasquie-Schiex, M.C.: On the acceptability of arguments in bipolar argumentation frameworks. In: European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty. pp. 378–389. Springer (2005)
8. Cyras, K., Rago, A., Albini, E., Baroni, P., Toni, F.: Argumentative xai: A survey. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence. pp. 4392–4399 (2021). https://doi.org/10.24963/ijcai.2021/600
9. Fan, C., Liu, J., Zhang, Y., Wong, E., Wei, D., Liu, S.: Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In: The Twelfth International Conference on Learning Representations (ICLR) (2024), https://openreview.net/forum?id=gn0mIhQGNM
10. Freedman, G., Dejl, A., Gorur, D., Yin, X., Rago, A., Toni, F.: Argumentative large language models for explainable and contestable claim verification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 14930–14939 (2025)
11. Gandikota, R., Materzyńska, J., Fiotto-Kaufman, J., Bau, D.: Erasing concepts from diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2426–2436 (2023). https://doi.org/10.1109/ICCV51070.2023.00230, https://openaccess.thecvf.com/content/ICCV2023/html/Gandikota_Erasing_Concepts_from_Diffusion_Models_ICCV_2023_paper.html
12. Gu, T., Zhou, Z., Huang, K., Liang, D., Wang, Y., Zhao, H., Yao, Y., Qiao, X., Wang, K., Yang, Y., Teng, Y., Qiao, Y., Wang, Y.: Mllmguard: A multi-dimensional safety evaluation suite for multimodal large language models. arXiv preprint arXiv:2406.07594 (2024)
13. Helff, L., Friedrich, F., Brack, M., Kersting, K., Schramowski, P.: Llavaguard: An open vlm-based framework for safeguarding vision datasets and models. arXiv preprint arXiv:2406.05113 (2024). https://doi.org/10.48550/arXiv.2406.05113
14. Hénin, C., Le Métayer, D.: Beyond explainability: Justifiability and contestability of algorithmic decision systems. AI & Society **37**(4), 1397–1410 (2022). https://doi.org/10.1007/s00146-021-01251-8
15. Lee, T., Yasunaga, M., Meng, C., Mai, Y., Park, J.S., Gupta, A., Zhang, Y., Narayanan, D., Teufel, H.B., Bellagente, M., Kang, M., Park, T., Leskovec, J.,

Zhu, J.Y., Fei-Fei, L., Wu, J., Ermon, S., Liang, P.: Holistic evaluation of text-to-image models. arXiv preprint arXiv:2311.04287 (2023), https://arxiv.org/abs/2311.04287
16. Leofante, F., Ayoobi, H., Dejl, A., Freedman, G., Gorur, D., Jiang, J., Paulino-Passos, G., Rago, A., Rapberger, A., Russo, F., Yin, X., Zhang, D., Toni, F.: Contestable ai needs computational argumentation. In: Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning. pp. 888–896 (2024). https://doi.org/10.24963/kr.2024/83
17. Li, Y., Du, Y., Lin, K., Chen, Z., Li, M., Jiang, Y., et al.: Evaluating object hallucination in large vision-language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 292–305 (2023)
18. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision – ECCV 2014. pp. 740–755. Springer (2014)
19. Liu, X., Zhu, Y., Gu, J., Lan, Y., Yang, C., Qiao, Y.: Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. arXiv preprint arXiv:2311.17600 (2023)
20. Lu, S., Wang, Z., Li, L., Liu, Y., Kong, A.W.K.: Mace: Mass concept erasure in diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6430–6440 (2024), https://openaccess.thecvf.com/content/CVPR2024/html/Lu_MACE_Mass_Concept_Erasure_in_Diffusion_Models_CVPR_2024_paper.html
21. Qiu, H., Chen, G., Zhang, M., Yang, M.: Safe text-to-image generation: Simply sanitize the prompt embedding. arXiv preprint arXiv:2411.10329 (2024), https://arxiv.org/abs/2411.10329
22. Qu, Y., Shen, X., He, X., Backes, M., Zannettou, S., Zhang, Y.: Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. In: Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security. pp. 3403–3417 (2023). https://doi.org/10.1145/3576915.3616679
23. Rago, A., Toni, F., Aurisicchio, M., Baroni, P.: Discontinuity-free decision support with quantitative argumentation debates. In: Proceedings of the 15th International Conference on Principles of Knowledge Representation and Reasoning. pp. 63–73 (2016)
24. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: Proceedings of the 38th International Conference on Machine Learning. pp. 8821–8831 (2021), https://proceedings.mlr.press/v139/ramesh21a.html
25. Rando, J., Paleka, D., Lindner, D., Heim, L., Tramèr, F.: Red-teaming the stable diffusion safety filter. arXiv preprint arXiv:2210.04610 (2022). https://doi.org/10.48550/arXiv.2210.04610
26. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022), https://openaccess.thecvf.com/content/CVPR2022/papers/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.pdf
27. Schramowski, P., Brack, M., Deiseroth, B., Kersting, K.: Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22522–22531 (2023)

28. Schramowski, P., Tauchmann, C., Kersting, K.: Can machines help us answering question 16 in datasheets, and in turn reflecting on inappropriate content? In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. pp. 135–145 (2022). https://doi.org/10.1145/3531146.3533192, https://doi.org/10.1145/3531146.3533192
29. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, A., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)
30. Tsai, Y.L., Hsu, C.Y., Xie, C., Lin, C.H., Chen, J.Y., Li, B., Chen, P.Y., Yu, C.M., Huang, C.Y.: Ring-a-bell! how reliable are concept removal methods for diffusion models? arXiv preprint arXiv:2310.10012 (2023), https://arxiv.org/abs/2310.10012
31. Turpin, M., Michael, J., Perez, E., Bowman, S.R.: Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. Advances in Neural Information Processing Systems **36**, 74952–74965 (2023)
32. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
33. Wang, Z., Hu, S., Zhao, S., Lin, X., Juefei-Xu, F., Li, Z., Han, L., Subramanyam, H., Chen, L., Chen, J., Jiang, N., Lyu, L., Ma, S., Metaxas, D.N., Jain, A.: Mllm-as-a-judge for image safety without human labeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14657–14666 (2025)
34. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E.H., Le, Q.V., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems **35**, 24824–24837 (2022)
35. Yang, Y., Gao, R., Yang, X., Zhong, J., Xu, Q.: Guardt2i: Defending text-to-image models from adversarial prompts. arXiv preprint arXiv:2403.01446 (2024), https://arxiv.org/abs/2403.01446
36. Yang, Y., Hui, B., Yuan, H., Gong, N.Z., Cao, Y.: Sneakyprompt: Jailbreaking text-to-image generative models. In: 2024 IEEE Symposium on Security and Privacy (SP). pp. 897–912 (2024), https://arxiv.org/abs/2305.12082
37. Zhang, Y., Huang, Y., Sun, Y., Liu, C., Zhao, Z., Fang, Z., Wang, Y., Chen, H., Yang, X., Wei, X., Su, H., Dong, Y., Zhu, J.: Multitrust: A comprehensive benchmark towards trustworthy multimodal large language models. arXiv preprint arXiv:2406.07057 (2024)
38. Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., Smola, A.: Multimodal chain-of-thought reasoning in language models. Transactions on Machine Learning Research (2024)
39. Zhou, K., Liu, C., Zhao, X., Compalas, A., Song, D., Wang, X.E.: Multimodal situational safety. arXiv preprint arXiv:2410.06172 (2024)