MIRAGE: Metadata-guided Image Retrieval and Answer Generation for E-commerce Troubleshooting

Rishav Sahay*, **Lavanya Tekumalla***, **Anoop Saladi**
Amazon
rishavsahayiit@gmail.com, lavanya.tekumalla@gmail.com
{saladias}@amazon.com

Abstract

Existing multimodal systems typically associate text and available images based on embedding similarity or simple co-location, but such approaches often fail to ensure that the linked image accurately depicts the specific product or component mentioned in a troubleshooting instruction. We introduce **MIRAGE**, a metadata-first paradigm that treats structured metadata, (not raw pixels), as a first-class modality for multimodal grounding. In MIRAGE, both text and images are projected through a shared semantic schema capturing product attributes, context, and visual aspects, enabling reasoning over interpretable attributes for troubleshooting rather than unstructured embeddings. MIRAGE comprises of three complementary modules: **M-Link** for schema-guided image–text linking, **M-Gen** for metadata-conditioned multimodal generation, and **M-Eval** for consistency evaluation in the same structured space. Experiments on large-scale enterprise e-commerce troubleshooting data across 10 product types on 100K text chunks and 35K images show that metadata-centric grounding achieves over 40 pp higher linking coverage of high-quality visual content and over 45 pp in linking and response quality than embedding-based baselines. MIRAGE demonstrates the potential of structured metadata in enabling scalable, fine-grained grounding in multimodal troubleshooting systems.

1 Introduction

Despite the rise of conversational AI and large language models (LLMs) [1, 11], most product troubleshooting experiences today remain largely text-based [16], especially in high-friction use cases like technical troubleshooting. Troubleshooting technical issues, especially for consumer electronics like headphones and mobile phones, often depends on the user's ability to identify and act on specific aspects such as buttons, icons, or ports on device. As a result, users frequently struggle to follow troubleshooting instructions like *"Open SIM tray by inserting ejection pin into the hole"* where visual cues could be much more intuitive (see fig 1).

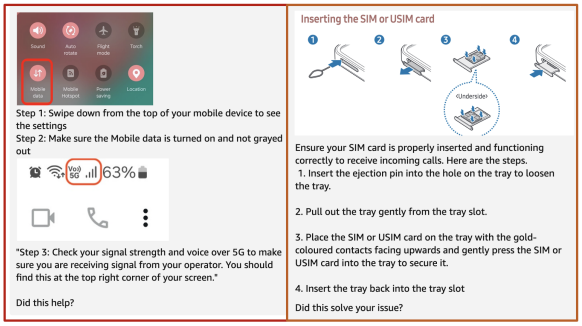

Figure 1: Illustrative example: Multimodal troubleshooting response to the query "Not receiving calls on my new phone."

Multimodal troubleshooting systems, which combine textual guidance from the underlying knowledge base (KB) with visual content, can dramatically improve user experience by offering more actionable help. The field has seen rapid progress, moving from general-purpose multimodal models ([19, 18, 6] to sophisticated Multimodal Retrieval-Augmented Generation (M-RAG) frameworks [15, 8] and benchmarks [10, 20]. However, despite these advancements, existing M-RAG systems are not well-suited for high-stakes troubleshooting workflows that require grounding in structured, brand-specific content and demand a systematic approach to ensuring factual consistency and task-oriented guidance.

Also, while rich visual content is widely available on brand support pages and e-commerce detail pages (DPs), it remains underutilized, rarely linked to troubleshooting KB chunks, resulting in poor visual coverage. A straightforward approach to establishing such links is embedding-based retrieval, where image and text embeddings are com-

*Equal contribution

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track), pages 764–776
March 25-27, 2026

pared in a shared representation space [14, 5] to link closest matching image to KB chunk. While scalable, these methods often fail to ensure factual alignment, sometimes surfacing the wrong product variant or irrelevant visuals. This highlights a key gap: the need for scalable, accurate methods to link, retrieve and generate visual content in troubleshooting workflows with factual consistency and systematic task-level evaluation framework to ensure detail alignment between text and images.

To address these limitations, we introduce **MIRAGE**, a scalable metadata-first framework built around a shared schema capturing product attributes (model, brand), troubleshooting context(customer query that is being solved), and fine-grained aspects (e.g., ports, buttons). This structured representation anchors three modules: **M-Link** links troubleshooting-relevant images with KB chunks through schema-guided matching and factual guardrails; **M-Gen** generates multimodal responses via a Retrieval-Augmented Generation (RAG) pipeline conditioned on *image metadata* rather than raw pixels for grounded reasoning; and **M-Eval** evaluates responses across four dimensions—Relevance, Attribute Alignment, Aspect Alignment, and Image Groundedness, providing a unified, metadata-consistent evaluation loop.

1.1 Contributions

- We introduce **MIRAGE**, a metadata-first paradigm that uses a shared (attribute, context, aspect) schema to treat metadata as a first-class modality anchoring text and image reasoning.
- We propose M-Link, a novel metadata-guided and guardrailed image-text linking algorithm that significantly improves image coverage and has higher factual alignment compared to direct embedding-based retrieval methods
- We propose M-Gen, a lightweight LLM-based retrieval-augmented multimodal response generation module that utilizes fine-grained image metadata instead of raw image inputs
- We propose M-Eval, the first evaluation framework tailored for multimodal troubleshooting RAG systems, with explicit guardrails on context and fine-grained domain specific attribute alignment to ensure factual accuracy.
- We evaluate MIRAGE across $\sim$100K KB chunks from 10 Product types showing 40 pp higher visual coverage and 45 pp improvement in image–text alignment over baselines.

These contributions present the first end-to-end, evaluation-driven framework for metadata-guided linking and generating multimodal troubleshooting responses in real-world enterprise settings.

2 Literature Survey

Multimodal Troubleshooting Systems Multimodal systems have advanced significantly across visual reasoning and generation tasks [2, 3, 19, 18, 6, 13]. Recent advancements in Multimodal Retrieval-Augmented Generation (M-RAG), including optimizations for industrial applications [15], multi-agent architectures [8, 17], and Multimodal-to-Multimodal Generation (M^2RAG) [10], have improved knowledge-grounded outputs. However, these models are not designed for structured, domain-specific troubleshooting workflows that rely on a technical knowledge base (KB) and demand precise visual and factual fidelity.

Image–Text Linking Existing systems for curating multimodal KBs rely on simple co-occurrence heuristics or general-purpose embedding-based retrieval methods [14, 5]. While useful, these techniques often fail in structured enterprise applications like troubleshooting, where fine-grained factual alignment between images and domain-specific text (e.g., procedural steps, device components) is critical. Weakly supervised approaches like AutoKnow [21] focus only on text-based entity linking. We argue that existing M-RAG indexing methods are insufficient for enterprise troubleshooting because of (1) poor coverage, as they ignore vast, unlinked visual content available outside the text vicinity, and (2) low factual precision, as they fail to leverage explicit metadata to enforce the necessary cross-modal factual alignment.

Multimodal Response Generation and Evaluation Alongside open-ended multimodal generation models [12, 7, 22], the field has seen the emergence of Multimodal Retrieval-Augmented Generation (M-RAG) frameworks [15, 8] and benchmarks [10, 20, 9, 4] that assess reasoning and factual grounding in open-domain settings. However, these efforts do not address the generation and evaluation of domain-specific multimodal responses where image aspect fidelity and fine-grained image–text alignment are essential. Our work bridges this gap through a metadata-guided framework that links, generates, and evaluates multimodal trou-

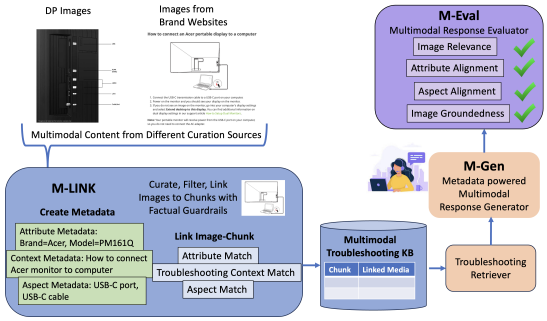

Figure 2: The MIRAGE framework for Multimodal Troubleshooting KB

bleshooting responses with explicit guardrails on context and fine-grained factual alignment.

3 MIRAGE Framework

Our framework comprises of (1) M-Link: Scalable Content curation and Metadata based linking of images with KB chunks (2) M-Gen: Generating multimodal responses leveraging image metadata (3) M-Eval: Evaluating multimodal responses for fine-grained factual consistency

3.1 M-Link: Cross-modal curation & Linking

We now describe our content curation, metadata generation and image-text linking algorithm.

3.1.1 Content Curation and Filtering

We curate relevant images from two key sources: (1) **Brand Support Webpages** often contain illustrative diagrams, annotated product shots, or step-by-step visuals for troubleshooting, but also include unrelated logos, icons, and promotional imagery, necessitating cleanup. (2) **Product Detail Pages (DPs)** contain high-quality product images including different angles, infographics, and port callouts; we prioritize images that display troubleshooting relevant aspects such as ports, buttons, indicator lights, error messages, and control panels that are useful for troubleshooting.

Each curated image is passed through a **LLM based relevance classifier**(Prompt A.2) to ensure the image is informative for troubleshooting.

3.1.2 Metadata Construction

We use a lightweight LLM (*claude-3-haiku*) to extract structured metadata from both KB chunks and images, creating a compact representation for matching

- **Product Attribute Metadata** ($\mathcal{P} = \{b, m\}$): Specifies product identity through its **brand** (b) and **model** (m) attributes, ensuring that image and text refer to the same product. Brand-level matches enable reuse, while model-level matches ensure precision.
- **Context Metadata** (μ): A concise description of the troubleshooting issue or scenario (e.g., "no sound from headphones") that could be answered using the image or text chunk.
- **Aspect Metadata** ($\mathcal{A}$): A set of visual elements or components (e.g., "HDMI port", "power button") present in the image or text chunk, serving as fine-grained visual anchors.

Metadata for chunk C_j and image I_i are derived as: (See Prompt A.3):

$$\texttt{LLM}(C_j) \rightarrow M_{C_j} = (\mathcal{P}_{C_j}, \mu_{C_j}, \mathcal{A}_{C_j})$$

$$\texttt{LLM}(I_i) \rightarrow M_{I_i} = (\mathcal{P}_{I_i}, \mu_{I_i}, \mathcal{A}_{I_i})$$

This abstraction enables efficient image–chunk linking without exhaustive comparisons.

3.1.3 Metadata-Guided Image Linking

A key challenge in multimodal troubleshooting is linking curated images I_i to KB chunks C_j. Direct embedding-based matching often misses fine-grained alignment signals such as product model, aspect, or intent. Prompting multimodal LLMs for every image-chunk pair is accurate but expensive to scale. We instead use structured metadata to guide and constrain linking.

Linking Rubric with Guardrails: To ensure factual alignment while linking candidate images to chunk C_j, we apply the following constraints:

1. **Product Attribute Metadata Match:** Require brand match and model-level consistency: $b_{C_j} = b_{I_i}$ and $(m_{C_j} = \emptyset \ \vee \ m_{C_j} = m_{I_i})$. Note that model agnostic chunks can match images of any model.
2. **Context Metadata Match:** Cosine similarity between embedded contexts must exceed threshold δ: $\texttt{sim}(\mu_{C_j}, \mu_{I_i}) \geq \delta$.
3. **Aspect Metadata Match:** Given aspect sets $\mathcal{A}_{C_j}$ and $\mathcal{A}_{I_i}$, a match is valid if $\exists a \in \mathcal{A}_{C_j}, a' \in \mathcal{A}_{I_i}$ such that $\texttt{sim}(a, a') \geq \epsilon$.

These constraints ensure relevance and factual grounding, enabling reuse of images across compatible products without introducing misleading content. Duplicate or near-identical images in the linked set for chunk C_j are removed using CLIP-based deduplication.

3.2 M-Gen: Multimodal Response Generation with Image Metadata

To generate actionable troubleshooting responses with a LLM, we leverage a RAG setup to get the right set of evidence chunks and visuals.

Retrieval of Relevant Chunks. Given a user query q, we retrieve the top-k most relevant KB chunks using an embedding-based retriever:

$$\text{Retrieve}(q) \rightarrow \{(C_1, M_{C_1}, M_{\mathcal{I}_{C_1}}), \ldots, (C_k, M_{C_k}, M_{\mathcal{I}_{C_k}})\}$$

where $\mathcal{I}_{C_k}$ is the list of all metadata of images linked to chunk C_k with attribute, context and aspect information.

LLM-Based Solution Generation We prompt an LLM with the user query q, the retrieved chunks $\{C_j\}$, and their associated image metadata to generate a set of most appropriate troubleshooting solutions $\{S_1, ..., S_n\}$ where each solution $S_i = \{s_1, ..., s_m\}$ is a sequence of troubleshooting steps. For each step, the LLM evaluates available image metadata and decides whether a visual aid is relevant. If so, it inserts a placeholder token (e.g., `<img_slot_i>`) after the step. See Prompt A.1 for a summary of the prompt. At runtime, image placeholders are replaced by their corresponding images. Examples of solution generation for customer queries is shown in table 1.

3.3 M-Eval: Multimodal Response Evaluation

We propose **M-Eval**, a task-specific evaluation framework to systematically assess the factual and visual quality of multimodal troubleshooting responses. Each generated solution consists of sequential textual steps $S = \{s_1, s_2, \ldots, s_n\}$, where each step s_i is optionally associated with a set of linked images $\mathcal{I}_i = \{I_{i1}, \ldots, I_{im}\}$.For every pair (s_i, I_{ij}), M-Eval computes:

Relevance (IR): Measures whether image I_{ij} provides meaningful visual support for the instruction in s_i rather than being generic or decorative. An LLM-as-a-Judge assigns a scalar score: $IR(s_i, I_{ij}) \in \{1, 2, 3, 4, 5\}$

Attribute Alignment ($AttA$): Checks product identity consistency by verifying that the product attribute metadata $\mathcal{P} = \{b, m\}$ (e.g., "Samsung","Q90 TV")) in step s_i matches that of image I_{ij}. The score is binary: $AttA(s_i, I_{ij}) = 1[b_{s_i} = b_{I_{ij}} \wedge (m_{s_i} = \emptyset \vee m_{s_i} = m_{I_{ij}})]$

Aspect Alignment ($AspA$): Checks whether any visual aspects in $\mathcal{A}$ are explicitly referenced in s_i (e.g., ports, buttons, indicators) are visible in I_{ij}. The score is ordinal: $AspA(s_i, I_{ij}) \in \{1, 2, 3, 4, 5\}$

Image Groundedness (IG): Ensures that I_{ij} originates from the same evidence chunk C_k that informed s_i, preventing cross-chunk image leakage (a frequent issue in multimodal retrieval-augmented generation). The groundedness score is binary: $IG(s_i, I_{ij}) = 1[C_{s_i} = C_{I_{ij}}]$

See Prompt in A.6 This structured formulation enables both automatic metric computation and LLM-as-a-Judge scoring for large-scale evaluation of multimodal troubleshooting responses.

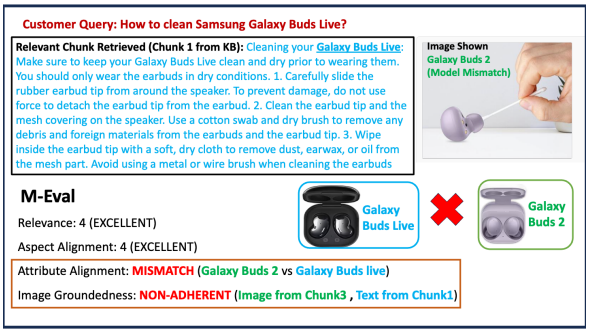

Figure 3: The customer has a query about cleaning *Samsung Galaxy Buds Live*. But the visual (to the right) shows Samsung Galaxy Buds2 being cleaned. While the the Relevance and Aspect alignment score is good, during response generation, the LLM picked the best chunk (chunk 1) to show response about Galaxy Buds Live, but picked an image linked to chunk 3, despite prompting to use images linked to solution chunks only. This chunk-image inconsistency is detected in the Image groundedness metric, while model mismatch is caught by the attribute alignment metric

4 Experiments

We evaluate MIRAGE on large-scale troubleshooting content from 10 product types, focusing on two core tasks: (1) image–text linking (M-link) (2) generated solution quality (M-Gen) with with our evalualtion framework (M-Eval).

4.1 Dataset and Setup

Knowledgebase and Multimodal content: We use a Knowledge Base with ∼100k text chunks derived from official support pages for 10 different product types (Eg: TVs, Laptops, Speakers,...) . The image pool contains about $35K$ curated visuals from brand websites (BrandImg) and $9K$ e-commerce detail pages (DPImg). Table 2 shows the chunk distribution across product types.

Product Query	Solution	Solution Image
Marshall Acton II Bluetooth speaker How do I turn on and use my Bluetooth speaker?	**Turn On and Use Your Marshall Acton II** • Hold the power lever to turn on. • Select Bluetooth mode using the Source button <*img_slot*>. • Pair with "Marshall Acton II" on your device. • Adjust Volume, Bass, and Treble. • Turn off by holding the power lever again.	
Samsung Television How do I charge and check the battery level of my solar-powered TV remote?	**Charging and Checking Battery Level of Samsung Solar-Powered TV Remote** • Check if charging is needed: If the LED shows for 5+ seconds, it's time to charge. • Use solar charging: Place the remote under light. • Or use a USB-C cable to charge <*img_slot*>. • Check battery: Press Home on the remote. • Go to Settings > All Settings > Power and Energy Saving > Available Remote Battery.	

Table 1: Fast Multimodal Solution Generation with **M-Gen** with Pre-linked Image-Metadata

4.2 Image–Text Linking Evaluation (M-Link)

Linking Coverage: We first evaluate our image linking coverage, which we define as the percentage of text chunks successfully tagged with at least one image. Our baseline, **ChunkLink**, relies solely on co-located brand images found within the text. We compare this against our two proposed methods: **BrandLink**, which uses metadata to link a broader set of brand images, and **DPBrandLink**, which further expands coverage by including DP images. Due to the confidential nature of the absolute figures, Table 2 reports the relative percentage improvement that BrandLink and DPBrandLink achieve over the ChunkLink baseline.

Linking Quality: We use LLM-as-a-judge (with *claude-4-sonnet*) to evaluate the quality of image-to-chunk links. For each image-chunk pair, LLM outputs a binary score (success: 1, failure: 0) for metrics (1) Relevance(IR) (2) Aspect Alignment(AscA) (3) Attributes Alignment(AttA) similar to metrics in sec 3.3. Prompt in app A.5.

For each metric, we compute the final score as percentage of linked images within the product type receiving a success (1). We evaluate our linking algorithm DPBrandLink against a baseline that uses CLIP embedding-based similarity(CLIPL) to link images to text and report the absolute improvement over CLIPL baseline in percentage points (pp) due to confidentiality reasons. See results in Tab 2.

4.3 Response Generation Evaluation (M-Gen)

We evaluate end-to-end multimodal solution quality on 1811 real user queries across 9 product types. For each query, relevant chunks are retrieved from the multimodal KB (∼100K chunks), and an LLM (*claude-3-haiku*) generates a solution consisting of text steps and optionally tags images as described in sec 3.2. To contextualize the improvements of M-Gen we compare against a strong embedding-based baseline CLIPL-Gen-Meta that uses CLIPL for image–chunk linking with metadata based multimodal responses. We evaluate two variants of M-Gen over this baseline. **M-Gen-Img**: takes raw images in the prompt and **M-Gen-Meta**: takes metadata of linked images for response generation.

Solution quality is evaluated with M-Eval using an LLM prompt (see App A.6) across four key dimensions: (1) Image Relevence (IR) (2) Aspect Alignment (AspA) (3) Attribute Alignment (AttA) (4) Image Groundedness (IG) as described in sec 3.3. We show improvement in percentage points of M-Gen-Img and M-Gen-Meta over CLIPL-Gen-Meta due to reasons of confidentiality in table 3. We also corroborated our findings with human evaluation on a subset of 500 queries (see tab 5).

4.4 Discussion of Results

Our experiments demonstrate that **metadata-guided DPBrandLink** substantially improves visual coverage. As shown in Table 2, coverage increases by approximately 40 pp on average when using DPBrandLink compared to using only proximity based chunk images (ChunkLink). Notably, linking from Brand websites contributes to more coverage compared to DP image based linking, suggesting that DP images are less suitable for troubleshooting scenarios. Also, as reported in Table 2, the linked images demonstrate high quality and relevance as compared to CLIPL based linked images, with an average gain of 43.22 pp in *IR*, and consistently high scores in *AscA* and *AttA* over clip baseline, indicating that the images are both relevant and factually matched to the product context.

Finally, our evaluation of end-to-end multimodal solution generation (M-Gen) is in table 3. *We note that **M-Gen-Meta** model is the best overall performer.* Both our models M-Gen-Img and M-Gen-

Table 2: **Dataset Overview and Linking Performance.** (1) Dataset statistics across product types; (2) Image Linking Coverage: Δ pp of BrandLink and DPBrandLink over ChunkLink: Leveraging Brand and DP images improves coverage significantly (3) Image Linking Quality: Δ pp improvement over CLIP based linking baseline: DPBrandLink significantly outperforms baseline

Category	(1) Dataset Statistics			(2) Image Linking Coverage		(3) Image Linking Quality		
	KB Chunks	DP Img	Brand Img	Δ BrandLink	Δ DPBrandLink	Rel	AscA	AttA
CELLULAR_PHONE	24812	387	10605	+32.71%	+34.57%	+53.55	+53.28	+21.71
HEADPHONES	13351	1778	7562	+40.86%	+45.58%	+49.06	+47.16	+36.33
NOTEBOOK_COMPUTER	12512	1595	11450	+34.35%	+36.57%	+49.55	+48.93	+21.50
SPEAKERS	7704	1042	1900	+25.88%	+37.85%	+48.95	+49.93	+28.52
MONITOR	2561	1558	1479	+21.44%	+29.84%	+50.16	+50.32	+30.80
VACUUM_CLEANER	2495	770	463	+8.29%	+37.43%	+19.95	+18.01	+9.39
TELEVISION	2394	1099	576	+48.08%	+57.27%	+36.42	+32.35	+15.24
REFRIGERATOR	947	364	288	+36.64%	+44.45%	+29.79	+29.34	+16.53
KEYBOARDS	427	90	242	+23.89%	+32.79%	+51.99	+47.90	+24.52
LAMP	90	0	2	+1.11%	+1.11%	+0	+0	+0

Table 3: Improvement over CLIPL-Gen (in Δ pp) for End-to-End M-Eval based Solution metrics: Bold indicates the better for each metric when at least one is positive. We note M-Gen-Image performs marginally better for relevance and aspect alignment, but M-Gen-Meta significantly outperforms in terms of Guardrail adherence.

Product Type	M-Gen-Image: With Images				M-Gen-Meta: With Image Metadata			
	IR	AscA	IG	AttA	IR	AscA	IG	AttA
HEADPHONES	**+46.3**	+47.2	-4.6	+30.0	+43.2	**+48.4**	**+2.8**	**+44.8**
WEARABLE_COMPUTER	**+24.7**	**+25.1**	-4.4	-12.4	+22.6	+24.0	**+2.4**	**+2.4**
CELLULAR_PHONE	**+44.6**	+40.3	-2.6	-18.8	+41.0	**+43.4**	**+7.2**	**+13.4**
SPEAKERS	**+26.7**	**+28.1**	-1.2	-10.2	+20.3	+23.6	**+1.3**	-15.9
NOTEBOOK_COMPUTER	**+34.0**	**+26.8**	-5.6	-3.8	+23.6	+24.7	-4.9	**+14.9**
KEYBOARDS	**+21.6**	**+15.0**	-2.2	-15.9	+16.4	+11.9	**+1.3**	**+3.9**
VACUUM_CLEANER	**+21.7**	+22.7	-6.1	+4.8	+20.3	**+27.1**	**+1.3**	**+24.3**
CAMERA_DIGITAL	**+3.4**	**+4.0**	-2.8	-24.1	+2.0	+2.5	-12.0	-16.8
ROBOTIC_VACUUM_CLEANER	+9.5	+12.1	-0.3	-45.3	**+16.2**	**+24.6**	**+0.8**	**+11.5**

Table 4: # tokens and latency: M-Gen-Image vs M-Gen-Meta

	M-Gen-Image		M-Gen-Meta	
	Tokens	Latency (s)	Tokens	Latency (s)
P50	3071	4.85	3070	5.02
P90	5218	7.26	4988	7.32
P99	8175	14.18	8338	13.42

Table 5: **Human Evaluation** between M-Gen variants across 6 product types on 500 queries: **M-Gen-Meta (A)** vs. M-Gen-Image (B). Labels: A>B = A better than B, B>A = B better than A, A=B = equally good, ALL_BAD = both poor.

Product Type (PT)	A>B	A=B	B>A	ALL_BAD
CELLULAR_PHONE	**0.5758**	0.1818	0.0909	0.1515
NOTEBOOK_COMPUTER	**0.6364**	0.2727	0.0909	–
MONITOR	**0.7500**	–	0.2500	–
VACUUM_CLEANER	**0.6000**	0.1000	0.3000	–
HEADPHONES	**0.6061**	0.1515	0.2424	–
WEARABLE_COMPUTER	**0.6207**	0.2759	0.0690	0.0345

Meta significantly outperform the CLIPL based baseline in terms of relevance metrics like *IR* and *AscA*. M-Gen-Img is often the most competitive in relevance metrics suggesting that using raw images may reduce information loss. However M-Gen-Img falters on critical grounding metrics *establishing* ***M-Gen-Meta*** *the winner*. The *M-Gen-Meta model is decisively stronger in* **AttA** *for attribute grounding and* **IG** that ensures linked images are surfaced from the right chunks. M-Gen-Meta leverages explicit metadata information to ensure factually grounded solutions that raw images might lack and it's superior performance is clearly corroborated by human evaluation on a subset of 500 queries in table 5.

Performance and Cost-Efficiency: Our system demonstrates strong practical utility for large-scale adoption. M-Gen's solution-generation takes sub-2s time-to-first-token and overall ∼5s (see table 4) while the offline linking of 100k chunks and 34k images is estimated to take 250 hours on a single AWS P4 node and can be faster with parallelization.

5 Conclusion

We present MIRAGE, a scalable framework that enriches troubleshooting workflows by (1) metadata-guided image-text linking with guardrails for factual alignment to improve coverage of visual content and (2) A novel framework for retrieval augmented multimodal solution generation and evaluation based on fine grained image-metadata for factual precision. Our evaluations across 100K KB chunks and 35K images show over 40 percentage point improvement in visual coverage and significant gains over baselines in multimodal solution quality and relevance across 9 product types.

6 Industrial Impact

MIRAGE is being integrated into a post-purchase chatbot for extending multimodal support in a large enterprise e-commerce workflow. Our text-based post-purchase troubleshooting chatbot, deployed across 8 marketplaces and 35 product types, re-

duced return rates by 6.5 bps and contact rates by 12.5%. However, 67% of failed troubleshooting cases stemmed from customer misinterpretation, highlighting the need for richer guidance. Motivated by this, we explored the role of visual cues and grounding for easier comprehension.

7 Limitations and Scope for Improvement

While MIRAGE has strong empirical performance of multimodal e-commerce troubleshooting, we describe some of the limitations of our study that present avenues for future work.

- **Domain Specificity.** Our evaluation focuses exclusively on the e-commerce troubleshooting domain. While the architectural principles are general, the specific metadata fields and the LLM prompts used for M-Link and M-Gen are highly tuned to this domain. Generalizing MIRAGE to a different domain (e.g., medical support or legal advice) would necessitate re-designing the metadata schema and extensive re-prompting/re-training.
- **Metadata Quality** "The core innovation of MIRAGE lies in the metadata-first paradigm, but this approach introduces a critical dependency. Errors in metadata (e.g., misidentifying the product model m or visual aspect A) directly lead to retrieving factually incorrect or irrelevant images, nullifying the benefit of the multi-modal approach. We are finetuning models that produce more accurate metadata.
- **LLM Hallucination and Safety.** The M-Gen component relies on a Lightweight LLM to synthesize the final answer. Like all LLMs, this model is susceptible to hallucination, where it may generate plausible but factually incorrect troubleshooting steps, especially when combining information from disparate text and image metadata. While we have text grounding metrics for this, we are making this more watertight.

8 Ethical Considerations

As MIRAGE is designed to provide actionable troubleshooting advice for consumer products, its deployment carries several important ethical considerations that must be addressed.

- **Safety and Factual Accuracy.** The primary ethical concern is the potential for generating unsafe or misleading instructions. Incorrect troubleshooting steps, particularly those involving electronics or physical components, could lead to user injury, product damage, or voiding warranties. We mitigate this through the use of guardrails in M-link and post-generation evaluation, but these must be rigorously maintained and audited to ensure safety.
- **Underlying LLM bias** Biases from the underlying LLM used for metadata and solution generation arising from training data distribution that might be more representative of certain geographies, product categories or ethnicity might impact troubleshooting solutions. The use of RAG architecture minimizes these biases to some extent in M-Gen through KB grounding. However a thorough study could lead to more insights on this aspect.
- **User Privacy.** In a real-world deployment, user queries contain sensitive information about their product usage, issues. Adherence to strict data governance and privacy policies is crucial to ensure that user interactions are protected.
- **Transparency of LLM Use.** It is ethically important to clearly communicate to the user that the troubleshooting advice is generated by an AI model (M-Gen) that combines text and image information through a metadata-guided process. This transparency manages user expectations regarding the source and certainty of the generated answer.

References

[1] Anthropic. 2025. Claude - anthropic. https://www.anthropic.com/claude/sonnet. Feb 2025.

[2] Ana Cláudia Akemi Matsuki de Faria, Felype de Castro Bastos, José Victor Nogueira Alves da Silva, Vitor Lopes Fabris, Valeska de Sousa Uchoa, Décio Gonçalves de Aguiar Neto, and Claudio Filipi Goncalves dos Santos. 2023. Visual question answering: A survey on techniques and common trends in recent literature. *arXiv*.

[3] Yang Deng, Wenqiang Lei, Wai Lam, and Tat-Seng Chua. 2023. A survey on proactive dialogue systems: Problems, methods, and prospects. *arXiv preprint arXiv:2305.02750*.

[4] Chaoyou Fu and 1 others. 2023. Mmbench: Is your multi-modal model an all-rounder? *arXiv preprint arXiv:2307.06281*.

[5] Junnan Li, Dongxu Li, Caiming Xiong, and Steven CH Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language under-

standing and generation. In *European Conference on Computer Vision (ECCV)*, pages 38–56. Springer.

[6] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

[7] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

[8] Pei Liu, Xin Liu, Yanlin Wang, Jian Zhang, Jiacheng Tu, and Jun Ma. 2025. HM-RAG: Hierarchical Multi-Agent Multimodal Retrieval Augmented Generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, New York, NY, USA. ACM.

[9] Yan Liu and 1 others. 2024. Imageeval: Benchmarking the factuality of image generation and editing. *arXiv preprint arXiv:2403.XXXX*.

[10] Zi-Ao Ma, Tian Lan, Rong-Cheng Tu, Yong Hu, Heyan Huang, and Xian-Ling Mao. 2024. Multi-modal Retrieval Augmented Multi-modal Generation: Datasets, Evaluation Metrics and Strong Baselines. *arXiv preprint*.

[11] OpenAI. 2024. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

[12] OpenAI. 2024. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

[13] OpenAI. 2024. Sora: Creating video from text. `https://openai.com/research/sora`.

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

[15] Monica Riedler and Stefan Langer. 2024. Beyond Text: Optimizing RAG with Multimodal Inputs for Industrial Applications. *arXiv preprint*.

[16] Rishav Sahay, Arihant Jain, Purav Aggarwal, and Anoop S V K K Saladi. 2025. Autokb: Automated creation of structured knowledge bases for domain-specific support.

[17] Rishav Sahay, Lavanya Sita Tekumalla, Purav Aggarwal, Arihant Jain, and Anoop Saladi. 2025. ASK: Aspects and retrieval based hybrid clarification in task oriented dialogue systems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 881–895, Vienna, Austria. Association for Computational Linguistics.

[18] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*.

[19] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v(ision). *arXiv preprint arXiv:2309.17421*.

[20] Qinhan Yu, Zhiyou Xiao, Binghui Li, Zhengren Wang, Chong Chen, and Wentao Zhang. 2025. MRAMG-Bench: A Comprehensive Benchmark for Advancing Multimodal Retrieval-Augmented Multimodal Generation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, New York, NY, USA. ACM.

[21] Honglei Zhang and 1 others. 2020. Autoknow: Self-driving knowledge collection for products of thousands of types. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery And Data Mining*, pages 2729–2739.

[22] Deyao Zhu and et al. 2023. Minigpt-4: Enhancing vision-language understanding with gpt-4-level capabilities. *arXiv preprint arXiv:2304.10592*.

A Appendix: Prompts

Prompt A.1:Solution Generation

Instruction: As a troubleshooting assistant, generate structured, multi-modal solutions for a user's product issue, using the provided text and image data.
Inputs: (1) **Query Details** (Query, Product Type, Brand, Model), (2) **Text Chunks XML** (contains <chunk url="..."> with text and optional <tagged_images>), (3) **Images XML** (contains <image id="..."> with metadata).
Key Tasks Constraints:
1. Synthesize Solutions: Create logical solutions by extracting and combining relevant steps from <text_chunks>.
2. Map Images to Steps: Use image metadata (purpose, description) to link each tagged image ID to the single, most relevant step derived from its source chunk.
3. Strict Image Uniqueness: CRITICAL - Assign each image ID to AT MOST ONE <relevant_images> tag in the entire output.
4. Cite Sources: Each <solution> must include a <cite_urls> tag listing the <url> of *all* source chunks that contributed steps.
5. Explain Reasoning: Detail all analysis, synthesis, and image mapping decisions within <thinking> tags *before* presenting any solutions.
Output Structure: (1) '<thinking>...</thinking>', followed by (2) one or more '<solution>' blocks. Each solution must contain a '<solution_heading>', '<solution_steps>' (with '<step>' and '<relevant_images>' tags), and '<cite_urls>'.

Prompt A.2:Image Troubleshooting Relevance

Instruction: As an AI assistant, evaluate a product image for its relevance in technical troubleshooting based on its clarity and visibility of key components.
Inputs: (1) **Image**, (2) **Product Type** (<product_type>), (3) **Product Details** (<product_details>), (4) **Related Context** (<context>), (5) **Image Link** (<image_link>).
Evaluation Criteria:
1. Clarity: Is the image clear and in focus?
2. Visibility of Key Components: Does the image clearly show important parts relevant to troubleshooting (e.g., ports, buttons, indicator lights, labels)?
3. Specificity: Does the image focus on specific components (Relevant), or is it a generic, blurry, or "in-box" view (Not Relevant)?
4. Context: Use the provided <context> to aid in judging relevance.
Output Requirements: Strict XML format only. Provide 2-3 lines of reasoning in <thinking> and the final verdict (Relevant / Not Relevant) in <relevance>.

```
<response>
  <thinking>
    [Your step-by-step reasoning (2-3 lines)]
  </thinking>
  <relevance>
  [Relevant / Not Relevant]
  </relevance>
</response>
```

Provide input using these placeholders:

```
<product_type>{pt}</product_type>
<product_details>{pd}</product_details>
<context>{context}</context>
<image_link>{url}</image_link>
```

Prompt A.3:Image Metadata Generation

Instruction: Analyze a product image to extract structured metadata: attributes (brand, model), visible aspects, and relevant troubleshooting queries (context).
Inputs: (1) **Image**, (2) **Product Type** (<product_type>), (3) **Product Details** (<product_details>), (4) **Related Context** (<context>).
Tasks:
1. Identify Attribute Metadata: Determine the **Brand** and **Model** of the product. Use visible logos, labels, or text in the image. If not visible, you may infer from the **Product Details** input.
2. Identify Visible Details/Aspects: Examine the image for specific, in-focus components, labels, indicators, ports, or states (e.g., 'HDMI 2 Port', 'Error code E:21'), not generic terms. Note details relevant to the <context>.
3. Generate Potential Queries (Context): Based *only* on the visible aspects and context, formulate troubleshooting queries (this corresponds to Context Metadata μ) that this specific image can help visually answer.
Output Requirements: Strict XML format only. Do not add any text outside the specified tags.

```
<thinking>
[Perform step-by-step reasoning]
</thinking>
<output>
  <brand>[Brand name, e.g., "Dell"]</brand>
  <model>[Model name, e.g., "XPS 13"]</model>
  <aspects>
    <aspect>[Specific visible detail]</aspect>
    ...
  </aspects>
  <queries>
    <query>[Product Issue/Usage query]</query>
    ...
  </queries>
</output>
```

Provide input using these placeholders:

```
<product_type>{pt}</product_type>
<product_details>{pd}</product_details>
<context>{context}</context>
```

Prompt A.4:Text Chunk Metadata Generation

Instruction: Analyze a product's troubleshooting text chunk to extract structured metadata: attributes (brand, model), mentioned aspects, and relevant queries (context).
Inputs: (1) **Product Type** (<product_type>), (2) **Product Details** (<product_details>), (3) **Text Chunk** (<text_chunk>).
Tasks:
1. Identify Attribute Metadata: Extract any **Brand** or **Model** names *explicitly mentioned* in the <text_chunk>.
2. Identify Mentioned Details/Aspects: Examine the <text_chunk> for specific, explicitly mentioned components, labels, indicators, ports, buttons, menu paths, error codes, actions, or states (e.g., 'HDMI port 1', 'Network Settings menu').
3. Generate Unique Queries (Context): Based *only* on the identified aspects, formulate unique troubleshooting queries (this corresponds to Context Metadata μ) that this specific <text_chunk> can directly answer.
4. Constraints: If no brand, model, or specific aspects are explicitly mentioned, leave the corresponding output tags empty (e.g., '<brand></brand>'). Do not predict aspects or queries if the text is uninformative.
Output Requirements: Strict XML format only. Do not add any text outside the specified tags.

```
<thinking>
[Perform step-by-step reasoning]
</thinking>
<output>
  <brand>[Brand name, e.g., "Sony"]</brand>
  <model>
  [Model name, e.g., "WH-1000XM5"]
  </model>
  <aspects>
   <aspect>[Specific mentioned detail]</aspect>
    ...
  </aspects>
  <queries>
    <query>[Product Issue/Usage query]</query>
    ...
  </queries>
</output>
```

Provide input using these placeholders:

```
<product_type>{pt}</product_type>
<product_details>{pd}</product_details>
<text_chunk>{context}</text_chunk>
```

Prompt A.5:Linked Images Evaluation

Instruction: As an AI assistant, evaluate if a product image is relevant to a specific troubleshooting text chunk and score its alignment.
Inputs: (1) **Image**, (2) **Text Chunk** (<chunk>), (3) **Local Issue** (<local_issue>), (4) **Global Issue** (<global_issue>), (5) **Product Context** (<product_context> with Brand, Model, Product Type).
Evaluation Criteria: You must score the image against *each* of the following (1=yes, 0=no):
1. Image Relevancy: Does the image visually clarify the action, state, or components described in the <chunk>?
2. Aspect Alignment: Does the image clearly show *specific aspects* (e.g., ports, buttons, error messages) relevant to the <local_issue>?
3. Attribute Alignment: Does the image visually match the **Product Context** (Brand, Model, Product Type)?
Final Verdict: The image is 'Relevant' *if and only if* all three scores are '1'.
Output Requirements: Strict XML format only.

```
<response>
  <image_relevancy_score>
  [0 or 1]
  </image_relevancy_score>
  <aspect_alignment_score>
  [0 or 1]
  </aspect_alignment_score>
  <attribute_alignment_score>
  [0 or 1]
  </attribute_alignment_score>
  <relevance>
  [Relevant / Not Relevant]
  </relevance>
  <reasoning>
  [Concise 1-3 sentence justification.]
  </reasoning>
</response>
```

Provide input using these placeholders:

```
<chunk>
{chunk}
</chunk>
<local_issue>{local_issue}</local_issue>
<global_issue>{global_issue}</global_issue>
<product_context>
Brand: {brand}
Model: {model}
Product Type: {pt}
</product_context>
```

Prompt A.6:Solution Evaluation Prompt

Instruction: Evaluate image-solution alignment in multimodal troubleshooting across five dimensions.
Inputs: (1) **Query Details** (Product, Query, Brand, Model), (2) **Troubleshooting Response** (LLM solution), (3) **Source Chunks XML**, (4) **Image Metadata XML**.
Evaluation Criteria:
1. Image Relevance: Evaluate if the image in <relevant_images> directly illustrates the <step> text. *Checks:* Visual-textual correspondence, instructional clarity, contextual accuracy. *Scale:* EXCELLENT (4) / GOOD (3) / FAIR (2) / POOR (1).
2. Aspect Alignment: Evaluate if named aspects (e.g., 'HDMI port', 'red light') in the step's text clearly and accurately align with the image. *Checks:* Component identification, state/condition match. *Scale:* EXCELLENT (4) / GOOD (3) / FAIR (2) / POOR (1).
3. Image Groundedness: Evaluate if the solution text (from <cite_urls>) and the image (from <tagged_images> in Source Chunks XML) originate from the *same* source chunk. *Scale:* ADHERENT / NON-ADHERENT.
4. Attribute Alignment: Evaluate if the image's metadata (Brand, Model) from the Image Metadata XML matches the customer's **Query Details**. *Checks:* Brand match, model match, visual compatibility. *Scale:* EXACT_MATCH / COMPATIBLE / MISMATCH.
5. Image Duplication (Overall): Evaluate the entire response for redundant images that show the same information without adding new value. *Scale:* NO_DUPLICATION / MINOR_DUPLICATION / SIGNIFICANT_DUPLICATION.

Algorithm 1 M-Link: Metadata-Guided Image–Text Linking with Guardrails (Chunk-Scoped)

Require: Curated images $\{I_i\}$ from Brand/DP sources; KB chunks $\{C_j\}$; thresholds δ (context similarity), ε (aspect edit distance)

Ensure: For each chunk C_j, a linked image set $\mathcal{I}_{C_j} = \{(I_i, M_{I_i})\}$

Metadata for chunks: For each C_j, obtain $M_{C_j} = (\mathcal{P}_{C_j}, \mu_{C_j}, \mathcal{A}_{C_j})$ with $\mathcal{P}_{C_j} = \{b_{C_j}, m_{C_j}\}$
Metadata for images: For each I_i, obtain $M_{I_i} = (\mathcal{P}_{I_i}, \mu_{I_i}, \mathcal{A}_{I_i})$ with $\mathcal{P}_{I_i} = \{b_{I_i}, m_{I_i}\}$
for each chunk C_j **do**
    $\mathcal{I}_{C_j} \leftarrow \emptyset$
    for each image I_i **do**
        Attribute guardrails: require $\mathcal{P}_{C_j}$ and $\mathcal{P}_{I_i}$ to match such that $b_{C_j} = b_{I_i}$ and $\big(m_{C_j} =$ **or** $m_{C_j} = m_{I_i}\big)$
        if attribute guardrails satisfied **then**
            Context match: $s \leftarrow \mathrm{sim}(\mu_{C_j}, \mu_{I_i})$; require $s \geq \delta$
            if $s \geq \delta$ **then**
                Aspect match: require $\exists\, a \in \mathcal{A}_{C_j},\ a' \in \mathcal{A}_{I_i}$ with $\mathrm{sim}(a, a') \geq \varepsilon$
                if aspect matched **then**
                    $\mathcal{I}_{C_j} \leftarrow \mathcal{I}_{C_j} \cup \{(I_i, M_{I_i})\}$
                end if
            end if
        end if
    end for
    Deduplication: on $\mathcal{I}_{C_j}$, remove near-duplicates (e.g., via CLIP or metadata clustering)
end for
return $\{\mathcal{I}_{C_j}\}_j$

B Appendix: Algorithms

Algorithm 2 M-Gen: Retrieval-Augmented Multimodal Response Generation using Image Metadata

Require: User query q; retriever $\mathrm{Retrieve}(\cdot)$; KB chunks $\{C_j\}$ with metadata $M_{C_j} = (\mathcal{P}_{C_j}, \mu_{C_j}, \mathcal{A}_{C_j})$; linked image sets $\{\mathcal{I}_{C_j}\}_j$ from M-Link where $\mathcal{I}_{C_j} = \{(I_i, M_{I_i})\}$ and $M_{I_i} = (\mathcal{P}_{I_i}, \mu_{I_i}, \mathcal{A}_{I_i})$

Ensure: Multimodal solution set $\{S_1, \ldots, S_n\}$ with associated image placeholders

Chunk retrieval: $\{C_{j_1}, \ldots, C_{j_k}\} \leftarrow \mathrm{Retrieve}(q)$ with corresponding $\{M_{C_{j_r}}\}$ and linked sets $\{\mathcal{I}_{C_{j_r}}\}$
Generation prompt: Construct $P_{\text{Gen}}(q, \{(C_{j_r}, M_{C_{j_r}}, \mathcal{I}_{C_{j_r}})\})$ capturing query intent, retrieved evidence, and image metadata
LLM response: Invoke an LLM with P_{Gen} to produce textual solutions $\{S_1, \ldots, S_n\}$ containing inline visual placeholders (`<img_slot>`)
Render-time replacement: Replace each placeholder with its linked image reference(s) from $\mathcal{I}_{C_j}$ based on provenance metadata; the generator consumes metadata only during selection
return $\{S_1, \ldots, S_n\}$

Algorithm 3 M-Eval: LLM-based Multimodal Response Evaluation

Require: Generated solutions $\{S_1, \ldots, S_n\}$; each $S = \{s_1, \ldots, s_m\}$ with linked images $\mathcal{I}_{s_\ell} = \{I_{\ell 1}, \ldots, I_{\ell r}\}$ and provenance chunks $\mathrm{src}(\cdot)$

Ensure: Percentage success for each metric: Image Relevance (IR), Attribute Alignment (AttA), Aspect Alignment (AspA), and Image Groundedness (IG)

for each step s_ℓ in all solutions **do**
  for each image $I_{\ell j} \in \mathcal{I}_{s_\ell}$ **do**
    Construct evaluation prompt $P_{\text{Eval}}(s_\ell, I_{\ell j})$ containing the step text, image metadata $(\mathcal{P}_I, \mu_I, \mathcal{A}_I)$, and LLM judging instructions for IR, AttA, AspA, and IG

$$(\text{IR}, \text{AttA}, \text{AspA}, \text{IG})_{s_\ell, I_{\ell j}} \leftarrow \text{LLM\_Judge}(P_{\text{Eval}})$$

    Convert categorical outputs to binary success indicators:

$$\text{IR}^* = [\text{IR} \geq 3], \quad \text{AspA}^* = [\text{AspA} \geq 3],$$

$$\text{AttA}^* = [\text{AttA} = \text{EXACT\_MATCH or COMPATIBLE}],$$

$$\text{IG}^* = [\text{IG} = \text{ADHERENT}]$$

    Record $(\text{IR}^*, \text{AttA}^*, \text{AspA}^*, \text{IG}^*)$ for this pair
  end for
end for
Compute final metric scores as percentage of successful pairs:

$$\text{Score}_{\text{IR}} = \frac{\sum \text{IR}^*}{N},$$

$$\text{Score}_{\text{AttA}} = \frac{\sum \text{AttA}^*}{N},$$

$$\text{Score}_{\text{AspA}} = \frac{\sum \text{AspA}^*}{N},$$

$$\text{Score}_{\text{IG}} = \frac{\sum \text{IG}^*}{N}$$

where N is the total number of $(s_\ell, I_{\ell j})$ pairs.

return $\{\text{Score}_{\text{IR}}, \text{Score}_{\text{AttA}}, \text{Score}_{\text{AspA}}, \text{Score}_{\text{IG}}\}$ as percentage of successes per metric.