

Pragmatic Semantic Communication of Scene Description for Blind and Low-Vision Users

Peng Wu

School of Control Science
and Engineering
Shandong University
202300171088@mail.sdu.edu.cn

Ruxiao Chen

Department of Civil and
Systems Engineering
Johns Hopkins University
rchen117@jh.edu

Shuaishuai Guo*

School of Control Science
and Engineering
Shandong University
shuaishuai_guo@sdu.edu.cn

Abstract

Artificial intelligence (AI) assistants for blind and low-vision (BLV) users deliver scene descriptions through a serial auditory channel, often followed by spoken user queries. A central challenge is that scene-description pipelines typically prioritize visually or corpus-salient content, whereas users’ follow-up questions often depend on low-salience attributes that are omitted during generation or compression. To diagnose where this mismatch arises, we introduce an encoder-agnostic framework that decomposes scene-description systems into four stages: visual encoding, length control, content selection, and audio scheduling. Applying this framework to 200 blind-user-captured VizWiz images and 500 questions reveals a *salience–relevance mismatch*: salience-based compression preserves identity information but discards low-salience attributes needed for attribute-bound questions. Improving the visual-encoding stage by replacing BLIP-base with stronger open-weight MLLMs raises semantic retention but does not eliminate this mismatch. As a lightweight repair, we propose CA-Keyword, which explicitly preserves attribute terms and recovers color-question answerability by 0.23 while producing descriptions 37% shorter than the sentence

itize visually or corpus-salient content, whereas users may ask about low-salience attributes that are omitted during generation or compression. We therefore frame scene description as a task-oriented semantic communication problem whose evaluation should reflect what the listener actually asks rather than similarity to reference captions. In pragmatic terms, the user’s question is a speech act, and an effective description should maximize question-relevant informational effect per unit of listening time, consistent with the relevance principles of Grice (Grice, 1975) and Sperber and Wilson (Sperber and Wilson, 1995).

To diagnose where question-relevant information is lost, we view a scene-description assistant as a four-stage pipeline (Fig. 1). A vision *encoder* turns a captured photograph into an initial description; a length-controlled *compressor* selects which words to keep at a given target length; a text-to-speech module renders the compressed description as audio, and a *scheduler* decides in what order successive messages are delivered over a serial audio channel; the *listener* (the BLV user) then issues a follow-up spoken question. We call *answerability* the property that the words actually delivered let the listener answer that question. Each stage instantiates one of the four cooperative maxims of Grice (1975): encoder faithfulness maps to *Quality*, compressor length to *Quantity*, compressor word selection to *Relation*, and scheduling of the interruptible audio channel to *Manner*. The maxim taxonomy is Grice’s; what is new here is (i) its mapping onto concrete AI pipeline stages, (ii) an *answerability*-based, question-conditioned operationalisation of each stage against real user questions rather than reference captions, and (iii) the diagnostic findings that this decomposition makes visible (below). The formal per-stage definitions are deferred to Section 3.

We evaluate on 200 blind-authored VizWiz photographs and 500 user questions across BLIP-base

1 Introduction

Deployed artificial intelligence (AI) assistants such as Be My AI and Seeing AI describe photographs aloud to blind and low-vision (BLV) users at scale (Gonzalez Penuela et al., 2024). Because these descriptions are delivered through a serial auditory channel and are often followed by spoken user questions, their practical value depends not only on visual fidelity or verbosity, but on whether the pipeline retains the information needed to answer those questions. This creates a central challenge. Scene-description systems typically prior-

*Corresponding author.

and the 2024 open-weight MLLMs Moondream2 and Qwen2-VL-2B, with Holm-corrected paired statistics; full setup in Section 4. Sections 5.1–5.4 report four stratum-level observations aligned to the four maxims, and Section 5.3.1 shows that upgrading the encoder to a modern MLLM shrinks only the *Quality* term while leaving the others intact. Section 6 presents CA-Keyword, a 30-line salience-override compressor that restores question-relevant attribute words. Section 7 discusses implications for evaluating machine scene description in accessibility settings.

Contributions. This paper makes four contributions.

- We propose an encoder-agnostic, four-stage decomposition of machine scene description as a task-oriented semantic communication pipeline whose stages align with Grice’s four cooperative maxims and are each measured against user questions rather than reference captions.
- We identify a *salience–relevance mismatch*: salience-based compression preserves identity-bound questions but degrades attribute-bound ones that depend on low-salience modifiers. Color *answerability* drops by 0.38 ($p < 0.001$, $d_z = 0.68$) and text-reading by 0.14, while other question types remain flat; this effect is hidden in aggregate results. Replacing BLIP-base with Moondream2 or Qwen2-VL-2B reduces the encoder-level *Quality* gap by about 0.16 in retention but does not improve downstream *answerability* or remove the mismatch.
- We introduce CA-Keyword, a 30-line salience-override compressor that restores attribute words at 37% shorter length than the sentence baseline and recovers color *answerability* by 0.23.

2 Related Work

Pragmatics and grounding. Computational pragmatics has re-emerged around Rational-Speech-Act models and pragmatically informed generation (Fried et al., 2023), typically evaluated on constructed reference games. We port the pragmatic view to a naturalistic, accessibility-critical setting in which the listener is a human with asymmetric visual access; common ground cannot be verified visually. We use *Relation gap* in the

Gricean sense to denote the discrepancy between what the description makes salient and what the listener’s next question actually needs, so a *Relation gap* of this kind is typically not repaired within the current turn.

Accessibility HCI and BLV captioning. Prior work has qualitatively documented BLV users’ preferences for image descriptions and their reactions to imperfect captions (Gonzalez Penuela et al., 2024; Adnin and Das, 2024; Xie et al., 2024). At the dataset level, VizWiz established visual question answering and image description from blind photographers (Gurari et al., 2018; Antol et al., 2015), and modern multimodal language models (Li et al., 2022; Liu et al., 2024; Gemini Team, Google, 2023; Wang et al., 2024) now power the deployed assistants; concurrent work has begun to characterise their systematic hallucination on object presence, attributes, and relations (Li et al., 2023; Guan et al., 2024; Sun et al., 2023). Closest to us, Gurari et al. (2020) analysed whether image captions contain the information needed to answer VizWiz questions in aggregate; we extend this by (i) stratifying answerability by question-attribute type, which surfaces the salience–relevance mismatch that aggregate metrics hide, and (ii) decomposing the loss across four pipeline stages aligned to Grice’s maxims.

Task-oriented semantic communication. Goal-oriented and task-oriented extensions of semantic communication (Gündüz et al., 2023) evaluate a transmitted message by its usefulness for a downstream task rather than by its bit-level fidelity. When the downstream task is a human follow-up question under a serial audio budget, the natural task metric is *answerability* (whether the listener can answer that question from what was actually delivered) rather than similarity to a reference caption. The audio channel itself is a single-server queue (Kleinrock, 1975) bounded by short-term memory, and the scheduling policy over that queue operationalises Grice’s *Manner* maxim under this physical constraint.

3 The Pragmatic Communication Framework

This section formalises the mapping sketched in the Introduction. The four-maxim taxonomy below is Grice’s; the mapping onto pipeline stages and its per-stage operationalisation as retention, latency,

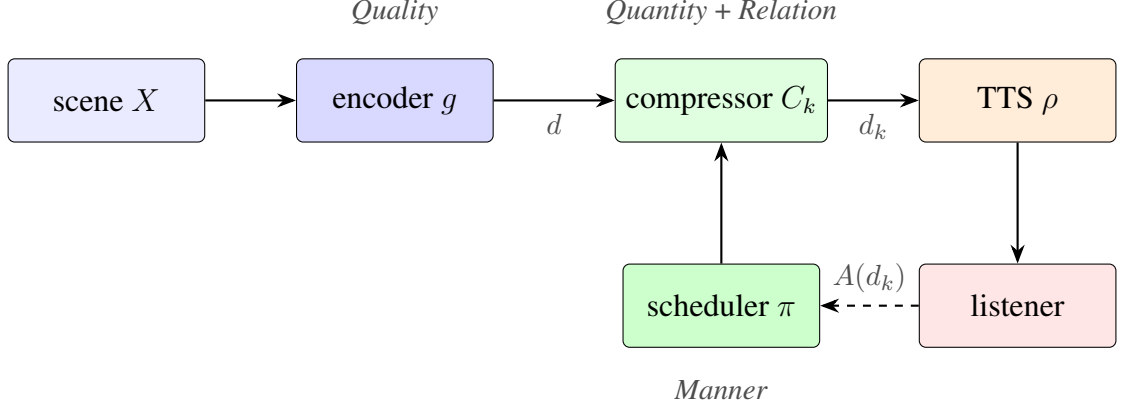


Figure 1: The machine scene-description pipeline studied in this paper. Each stage instantiates one Gricean cooperative maxim, and each is measured separately in Sections 5.1–5.4.

and answerability against real user questions are our contribution.

Grice (1975) decomposes cooperative communication into four maxims. *Quantity* asks the speaker to be as informative as required and no more; *Quality* to say only what is believed true; *Relation* to be relevant to the current purpose; and *Manner* to be perspicuous, orderly, brief, and unambiguous. Relevance theory (Sperber and Wilson, 1995) unifies *Quantity* and *Relation* into a single ratio of cognitive effect over processing effort, which is particularly apt when the “processing effort” has a physical unit: for speech, it is time.

The four stages of the pipeline in Fig. 1 map onto the four maxims as follows. *Quality* is enacted by the encoder g : a caption whose content is not entailed by the scene, so $D_{\text{gen}} > 0$, violates *Quality*, and semantic retention against held-out human descriptions is the operationalisation used here. *Quantity* is enacted by the compressor’s length choice $|d_k|$; under a fixed auditory time budget T , over-informativeness delays the next communicative act and under-informativeness starves the follow-up. *Relation* is enacted by the compressor’s word-selection policy: given the user’s question q , the retained words should be those whose answers are entailed by d , so the task-grounded operationalisation is answerability $A(d_k) = \text{Acc}_{\text{VizWiz}}(\text{QA}(d_k, q))$. *Manner* is enacted by the scheduler π : on a serial audio channel, an urgent message ordered after a routine one waits out the residual duration of the routine one, regardless of priority.

Under this mapping, the total delivered relevance

loss decomposes as

$$\underbrace{1 - S(C_k(\hat{g}(X)))}_{\text{delivered loss}} = \underbrace{D_{\text{gen}}(g)}_{\text{Quality}} + \underbrace{[1 - S(C_k(g(X)))]}_{\text{Quantity + Relation}} + \underbrace{\Delta_{\text{Manner}}(\pi)}_{\text{Manner}}. \quad (1)$$

where $\Delta_{\text{Manner}}(\pi)$ scales the first two terms by the fraction of the audio budget consumed by the scheduler π . Equation (1) indicates where a relevance gap is located and whether adding words at one stage can compensate a gap at another. Empirically, inter-annotator agreement on the same photographs gives a natural reference band $S_{\text{human}} = 0.638$. Retention losses inside this band are indistinguishable from substituting one blind-authored describer for another, and losses outside it are the maximum-level effects quantified in Sections 5.1–5.4.

4 Task, Data, Metrics, Statistics

Corpus. We use 200 photographs from VizWiz-Captions (Gurari et al., 2018, 2020), each with five blind-authored human captions, and 500 image-question-answer triples from the intersection with VizWiz-VQA (Antol et al., 2015). Two captions per image are held out as reference R ; compression levels are constructed from the remaining captions only, precluding self-matching. Of these 500 pairs, 121 also carry MLLM-generated captions and are used for the encoder comparison in Section 5.3.1. VizWiz questions were originally submitted together with the photograph rather than after a spoken description; we use them as a proxy

for the information a BLV user seeks about their own photograph, matching the distribution of post-description follow-ups in deployment.

Compression levels. Four operating points along the length axis are constructed: **Union**, the union of the three non-reference captions at 32.7 words on average; **LongCap**, the longest single source caption at 15.2 words; **ShortCap**, the shortest single source caption at 9.8 words; and **Keyword**, the top-five TF-IDF keywords at 5.0 words on average. These four levels are not proposed as a compression algorithm; they serve as representative operating points.

Encoders. Human union captions realize a “high-Quality speaker” baseline. BLIP-base (Li et al., 2022) realizes a machine speaker; we generate BLIP-greedy at 8.9 words with greedy decoding and BLIP-beam at 25.1 words with beam search plus a length constraint, to separate *Quality* effects from *Quantity* ones.

Metrics. With E a sentence embedder, retention $S(d_k) = \cos(E(d_k), \bar{E}(R))$. Answerability $A(d_k)$ is official VizWiz accuracy of a QA listener reading only d_k plus the user question. DistilBERT-SQuAD is the main listener and RoBERTa-SQuAD2 is the robustness listener.

Statistics. Bootstrap paired 95% confidence intervals with 5,000 resamples are used throughout, together with Holm–Bonferroni-corrected paired significance across each reported contrast family and paired effect sizes $d_z = \bar{\Delta}/\hat{\sigma}_\Delta$. All statistical tests are two-sided at $\alpha = 0.05$.

5 Findings Across the Four Gricean Maxims

5.1 *Quantity*: aggregate answerability is robust to threefold shortening

If the *Quantity* maxim were binding at conversational bandwidths, aggressive shortening should reduce a listener’s ability to answer the user’s real question. It does not. Table 1 reports DistilBERT accuracy on the 500 questions: Union 0.451, LongCap 0.417, ShortCap 0.421, Keyword 0.398. Six pairwise bootstrap contrasts give a single nominally significant difference between Union and Keyword at raw $p = 0.026$, and Holm correction across the family removes it, with the smallest adjusted $p = 0.156$ and all $|d_z| < 0.11$. A second listener

Type	N	Union	ShortCap	Keyword	Δ
Aggregate	500	0.451	0.421	0.398	+0.053
attribute-bound, relevance in a low-salience modifier					
color	62	0.645	0.366	0.263	+0.382
text-reading	32	0.490	0.490	0.354	+0.135
identity-bound, relevance in the head noun					
identification	190	0.491	0.551	0.495	−0.004
other	154	0.413	0.361	0.413	0.000
yes/no	57	0.193	0.199	0.222	−0.029

Table 1: Answerability, aggregate and stratified by question type. Δ is the Union-to-Keyword change. Attribute-bound strata collapse while identity-bound strata stay flat; Holm-corrected significance in the text.

(RoBERTa-SQuAD2) preserves the ordering, indicating the result is a property of the descriptions rather than of the listener.

The absolute answerability of 0.451 at Union is informative in itself. The binding constraint on this task is not the length of the description but its coverage of the words the user is likely to ask about. Adding more words to a description generated without knowledge of the question q cannot fill a coverage gap, whereas a question-conditioned selector can, as shown in Section 6.

5.2 *Relation*: a salience–relevance mismatch on attribute-bound questions

The aggregate result of Section 5.1 does not hold uniformly across question types, and the pattern of exceptions is systematic rather than idiosyncratic. Stratifying the 500 questions by a lightweight heuristic splits them into two regimes in Table 1 and Fig. 2. Identity-bound questions, namely identification, yes/no, and the residual “other” class, ask about the head noun of the scene, and their answerability is flat across a threefold compression, with $|\Delta| \leq 0.03$. Attribute-bound questions ask instead about a modifier of that noun, and they collapse. Color drops by 0.382 from Union to Keyword at Holm-adjusted $p < 0.001$ and $d_z = 0.68$, and text-reading by 0.135. Counting and spatial questions move in the same direction but are too sparse in this subset, with $n \leq 3$, to test.

We read this as a *salience–relevance mismatch*. A length-controlled compressor selects tokens by corpus salience such as frequency or TF-IDF, while answerability depends on task relevance, i.e. whether the retained tokens include the one the question asks about. For an identity-bound question the two coincide, since the head noun is both the most salient token and the relevant one, so

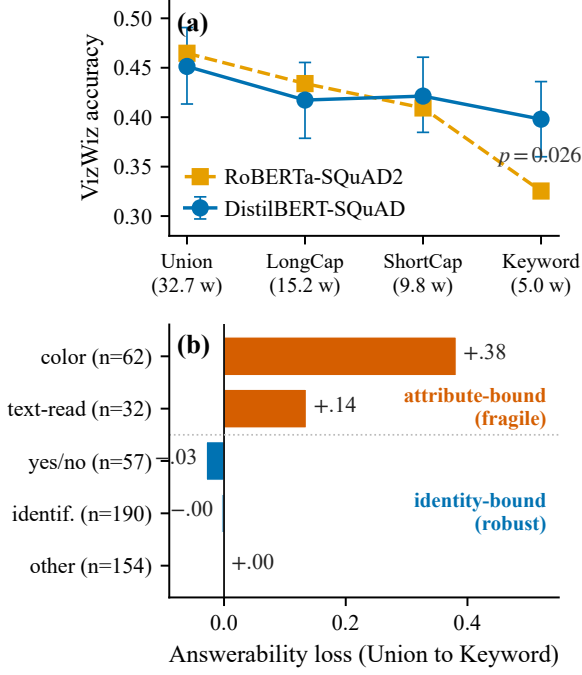


Figure 2: (a) Aggregate answerability across four length levels for both QA listeners. (b) Attribute-bound strata collapse while identity-bound strata are flat. Bars show the Union-to-KeyWord change over 500 questions.

compression is harmless. For an attribute-bound question they diverge, since the relevant token is a low-frequency modifier such as a color adjective, an OCR string, or a count, which salience-based selection discards first. The severity of the drop therefore scales with how far a question’s answer sits from the head noun, which is why color, the most modifier-dependent stratum, shows the largest gap and identification the smallest.

The mechanism is visible in a single example. For the user question “Can you tell me what color these bananas are?”, the ShortCap sentence reads “A bunch of yellow bananas with brown spots”, which keeps the relevant modifier. The TF-IDF keyword code, in contrast, reads “bananas, finger, blocking, banana, photo”: every content word survives except the one the user asked about. Compressors for BLV-style visual question answering should therefore be reported per question type (Levinson, 2000) with Holm-corrected per-stratum tests.

5.3 Quality: encoder distortion and the length lever

We regenerate captions for the same 200 photos using BLIP-base (Li et al., 2022). BLIP-greedy with greedy decoding at 8.9 words scores 0.543 in reten-

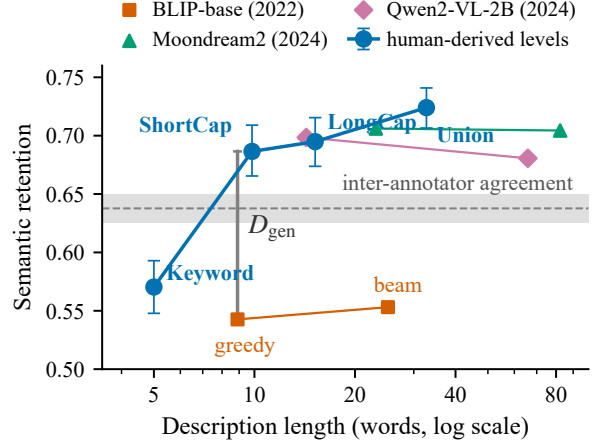


Figure 3: Semantic retention vs. code length (log scale). BLIP-base (squares) sits below the length-matched human baseline (circles); the *Quality* gap D_{gen} (vertical bar) exceeds the entire compressor-length range. The 2024 MLLMs (Moondream2, triangles; Qwen2-VL, diamonds) close most of D_{gen} but their long variants gain no retention. Mean $\pm$ 95% bootstrap CI over 200 photographs.

tion, which is below the inter-annotator band and 0.14 below the length-matched human ShortCap in Fig. 3. The resulting *Quality* gap D_{gen} exceeds the entire Union-to-KeyWord range. Tripling the output length with BLIP-beam at 25.1 words recovers only 0.010, indicating that additional words from a *Quality*-limited encoder do not compensate the residual *Quality* gap.

5.3.1 Do modern multimodal LLMs close D_{gen} ?

A natural worry is that our diagnosis is specific to BLIP-base from 2022. To test this we regenerate captions on the identical 200-photo set with two 2024 open-weight multimodal language models, Moondream2 1.9 B (Korrapati, 2024) and Qwen2-VL-2B-Instruct (Wang et al., 2024), and on the 121 image-question pairs whose photograph belongs to our subset, we also push their captions through the same extractive listener used in Section 5.2. Because MLLM outputs share no source with human captions, no leakage concern arises and the human operating points in Table 2 use all five captions per image, yielding a 44.6-word absolute Union upper bound rather than the 32.7-word non-reference Union of Table 1. Table 2 summarises the six-way comparison.

Three observations follow. **First**, both MLLMs raise retention by ≈ 0.16 over BLIP-base, so D_{gen} shrinks substantially; their long variants gain no fur-

Context supplied to listener	words	ret.	Acc.	lat.(s)
Human Union	44.6	–	0.556	–
Human ShortCap	6.9	0.616	0.477	–
Human Keyword	4.9	0.407	0.468	–
BLIP-greedy	8.9	0.543	–	–
BLIP-beam	25.1	0.553	–	–
Moondream2-short	23.1	0.706	0.466	0.52
Moondream2-long	82.3	0.704	0.366	1.64
Qwen2-VL-short	14.3	0.698	0.521	0.76
Qwen2-VL-long	65.9	0.681	0.433	2.52

Table 2: Modern MLLMs on the 200-photo subset for $N=200$ retention and on 121 held-out VizWiz-VQA pairs for $N=121$ accuracy; latency per image on an RTX 4060 laptop GPU at fp16. MLLMs shrink D_{gen} but not answerability; both long variants are significantly worse listeners than the human union.

ther retention, so length is not a lever for MLLMs either. **Second**, the retention gain does not carry over to answerability. Qwen2-VL-short scores 0.521 against Human Union’s 0.556 at $p = 0.11$, not significant, and both long variants are significantly worse than Human Union. The salience–relevance mismatch of Section 5.2 survives the encoder upgrade: on the color stratum the Union-to-MLLM gap is 0.13–0.17, comparable to the compressor-induced Union-to-Keyword gap, while identity-bound strata remain flat. **Third**, on latency, a Qwen2-VL-long encode at 2.52 s is already a quarter of the 10.1 s residual head-of-line window in Fig. 4, so an MLLM-front emergency pipeline must either pre-encode offline or pay this cost on top of the residual-blocking term.

This has an interpretive consequence for the framework. The *Quality* term does shrink monotonically as encoders improve, consistent with H_1 ; but the *Quantity+Relation* term (Sec. 5.1, Sec. 5.2) is orthogonal to encoder capacity, and the *Manner* term (Sec. 5.4) is a channel constraint independent of what the encoder does. Substituting BLIP-base with LLaVA-1.5, GPT-4V, or any future large model tightens D_{gen} but does not collapse the decomposition to a single stage.

5.4 Manner: priority scheduling under variable-length utterances

A standard treatment of delayed urgent messages is priority queueing: urgent items are placed at the head of the queue. On a serial audio channel the effect of this policy is bounded by the residual duration of the utterance currently being spoken.

We simulate 30-minute sessions with routine messages arriving Poisson at $\lambda_r=15$ s and urgent

Policy	Urg. delay (s)	Info del.	Drop
Detailed-FIFO	6.52 ± 1.98	2250 ± 129	34.1%
Concise-FIFO	0.83 ± 0.43	918 ± 77	0.2%
Priority-only	4.56 ± 1.36	2233 ± 131	34.7%
Adaptive [ours]	0.00 ± 0.00	1918 ± 133	27.1%

Table 3: Scheduling policies over 100 runs per policy (mean $\pm$ std). Only Adaptive reaches zero mean urgent delay, exact by construction, at a 15% information cost from preemption.

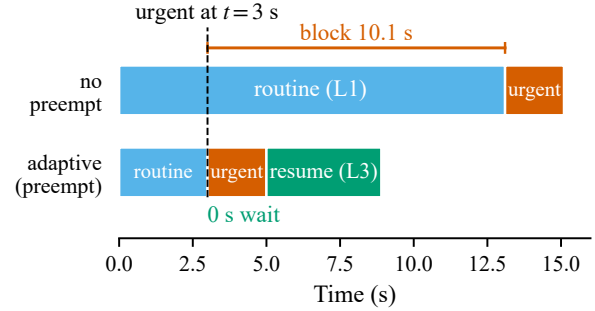


Figure 4: Head-of-line blocking on a serial audio channel. Top: a warning arriving mid-utterance waits 10.1 s under any non-preemptive policy. Bottom: preemption closes the gap at the information cost in Table 3.

events Poisson at $\lambda_u=120$ s. Preempted messages forfeit their information, so preemption is charged an explicit cost. Table 3 compares four policies. The key comparison is Priority-only versus Detailed-FIFO: Priority-only lowers mean urgent delay from 6.5 to 4.6 s, but the maximum urgent delay increases to 17.1 s. Thirteen-second Union codewords block the queue for their own duration regardless of priority. Only Adaptive, which switches an urgent event to the shortest code and permits preemption of the message currently being spoken, reaches zero mean delay, at an information cost of 15% from preemption (Fig. 4).

A sensitivity sweep over 12 combinations of arrival rates and TTS rates confirms this pattern. Adaptive delivers lower urgent delay than Priority-only at every grid point, and also a lower routine drop rate. Head-of-line blocking is present whenever message duration is comparable to the inter-arrival interval, which is the regime of realistic serial-audio assistants. On a variable-length audio channel, urgency cannot be reduced to a priority tag. Any scheduling policy that omits an explicit preemption-cost term will overestimate its own delay guarantee by the fraction of a routine message’s residual duration.

6 Pragmatic Repair: Saliency Override

The saliency–relevance mismatch of Section 5.2 suggests a direct repair: override corpus saliency for the low-frequency token classes that carry attribute relevance, so that they survive compression even though frequency-based selection would discard them. **CA-Keyword** is a deliberately minimal, non-learned instance of this idea, implemented in fewer than 30 lines. It extracts the attribute classes we can enumerate with a closed list of color and number tokens from the source, and appends up to two of them to the Keyword code at 6.2 words on average. In the banana example, the code becomes “bananas, finger, blocking, banana, photo, yellow, brown”.

Paired against Keyword on the 500 questions, CA-Keyword improves aggregate answerability from 0.398 to 0.427 at raw $p = 0.049$, and on the attribute-bound color stratum from 0.263 to 0.495 at raw $p = 0.006$ and Holm-adjusted $p = 0.013$, while no identity-bound class is harmed. At 6.2 words on average, CA-Keyword also exceeds the answerability of the 9.8-word ShortCap sentence at shorter length, so the Sperber–Wilson relevance ratio, effect divided by effort, increases. Because the repair acts on the description text rather than on the encoder, it is encoder-agnostic: it applies unchanged to a BLIP caption or to an MLLM caption that has grounded the attribute but let the compressor drop it. CA-Keyword covers only the two attribute classes expressible as a closed lexicon; text/OCR and spatial relations, which Section 5.2 shows are also attribute-bound, require a learned question-conditioned selector and are left as the natural extension of the same saliency-override principle.

7 Discussion and Conclusion

Discussion. Our experiments support the central claim that practical usefulness depends on retaining question-relevant information rather than preserving overall caption content. Compressing Human Union to Keyword reduces aggregate *answerability* only from 0.451 to 0.398, with no significant difference after Holm correction, but stratified results show a 0.38 drop for color questions and a 0.14 drop for text-reading, while identity-bound question types remain approximately flat. This contrast reveals a *saliency–relevance mismatch*: saliency-based compression preserves head nouns but removes the low-saliency modifiers needed for

attribute-bound questions, a failure hidden by aggregate evaluation. Moondream2 and Qwen2-VL-2B improve semantic retention by approximately 0.16 over BLIP-base without eliminating the mismatch, showing that stronger visual encoding alone is insufficient. CA-Keyword raises color *answerability* from 0.263 to 0.495 while remaining shorter than the sentence baseline, confirming that the diagnosed failure can be addressed through targeted content selection.

Future work. Absolute answerability values here rest on extractive QA listeners and therefore support between-condition comparison rather than human-level scores; a validation study with BLV participants over an accessible keyboard-and-TTS interface is the natural next step, and a generative listener such as LLaMA-Omni (Fang et al., 2024) would decouple the long-caption drop from the context-length sensitivity of extractive QA. The MLLM comparison uses 2 B-scale checkpoints, and whether 7 B+ or proprietary encoders such as GPT-4V or Gemini would also match rather than exceed human-union answerability is left for that same study.

Conclusion. We introduced an encoder-agnostic, four-stage diagnostic framework for evaluating BLV scene-description systems against real user questions rather than reference captions. Applying the framework to VizWiz revealed a *saliency–relevance mismatch*: compression preserves identity-bound information but discards low-saliency attributes needed for attribute-bound questions, a failure hidden by aggregate *answerability*. Stronger MLLM encoders reduce the encoder-level *Quality* gap but do not eliminate this mismatch. CA-Keyword provides a lightweight proof-of-concept repair by restoring attribute terms at shorter length than the sentence baseline, while the scheduling analysis shows that urgent delivery over a serial speech channel requires preemption rather than priority alone.

Ethics. All data are public and de-identified with no new subjects, and all models are open-weight checkpoints run locally.

Acknowledgements

This work was supported in part by Shandong University Research Project on Laboratory Development and Management under Grant sy20261301,

and in part by Shandong Provincial Natural Science Foundation under Grant ZR2025MS983. We thank the anonymous reviewers for their constructive feedback.

References

- Rudaiba Adnin and Maitraye Das. 2024. “I look at it as the king of knowledge”: How blind people use and understand generative AI tools. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS)*.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
- Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2024. LLaMA-Omni: Seamless speech interaction with large language models. *arXiv preprint arXiv:2409.06666*.
- Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. 2023. Pragmatics in language grounding: Phenomena, tasks, and modeling approaches. In *Findings of the Association for Computational Linguistics: EMNLP*.
- Gemini Team, Google. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Ricardo E Gonzalez Penuela, Jazmin Collins, Cynthia L Bennett, and Shiri Azenkot. 2024. Investigating use cases of ai-powered scene description applications for blind and low vision people. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.
- H Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry L Morgan, editors, *Syntax and Semantics, Vol. 3: Speech Acts*, pages 41–58. Academic Press.
- Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Deniz Gündüz, Zhijin Qin, Inaki Estella Aguerri, Harpreet S Dhillon, Zhaohui Yang, Aylin Yener, Kai Kit Wong, and Chan-Byoung Chae. 2023. Beyond transmitting bits: Context, semantics, and task-oriented communications. *IEEE Journal on Selected Areas in Communications*, 41(1):5–41.
- Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigam. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. Captioning images taken by people who are blind. In *European Conference on Computer Vision (ECCV)*.
- Leonard Kleinrock. 1975. *Queueing Systems, Volume 1: Theory*. Wiley-Interscience.
- Vikhyat Korrapati. 2024. Moondream: a small, fast, capable vision-language model. <https://github.com/vikhyat/moondream>. Version 2, model card on Hugging Face.
- Stephen C Levinson. 2000. *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. MIT Press.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning (ICML)*.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Dan Sperber and Deirdre Wilson. 1995. *Relevance: Communication and Cognition*, 2 edition. Blackwell.
- Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2023. Aligning large multimodal models with factually augmented RLHF. *arXiv preprint arXiv:2309.14525*.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-VL: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Jingyi Xie, Rui Yu, He Zhang, Sooyeon Lee, Syed Masum Billah, and John M Carroll. 2024. Emerging practices for large multimodal model (LMM) assistance for people with visual impairments: Implications for design. *arXiv preprint arXiv:2407.08882*.