Supplementary Material of RobustVisRAG: Causality-Aware Vision-Based Retrieval-Augmented Generation under Visual Degradations

I-Hsiang Chen[1] Yu-Wei Liu[1] Tse-Yu Wu[1] Yu-Chien Chiang[1]
Jen-Chieh Yang[1] Wei-Ting Chen[2]

[1]National Taiwan University [2]Microsoft

1. Distortion-VisRAG Dataset Details

To evaluate how visual distortions affect retrieval accuracy, generation reliability, and end-to-end stability in vision-based RAG systems, we construct the Distortion-VisRAG (DVisRAG) benchmark by extending the original VisRAG [20] dataset with diverse synthetic and real-world degradations. In addition, we include an out-of-domain (OOD) evaluation set based on RVL-CDIP [3] to assess generalization beyond the VisRAG distribution and to stress-test robustness under unseen document styles and degradation patterns. Summary statistics of all datasets are provided in Table 1. In the following sections, we describe the three components of DVisRAG: the Synthetic Degradation Dataset, the Real-World Degradation Dataset, and the OOD Zero-Shot Testing Set.

1.1. Synthetic Degradation Pipeline

To simulate diverse and challenging visual scenarios, we augment all VisRAG source datasets with 12 types of synthetic degradations generated following the procedure of [1]. These degradations include: exposure change, contrast variation, elastic transform, pixelation, JPEG compression, Gaussian noise, impulse noise, shot noise, motion blur, defocus blur, glass blur, and zoom blur. Each degradation type is applied at five severity levels, enabling fine-grained control over distortion intensity and improving the robustness of RobustVisRAG across varying image qualities. We also retain the original clean images in the training set, allowing the model to preserve high fidelity on undistorted inputs while learning robust degradation-invariant representations. Sample degraded images are shown in Figure 1, illustrating that these synthetic distortions effectively cover a wide range of realistic and challenging visual conditions encountered in document understanding tasks.

Table 1. **Statistics of the DVisRAG benchmark.** DVisRAG contains a total of 367,608 question–document (Q–D) pairs, including 362,110 training samples and 3,607 synthetic-degradation test samples, while the real-world degradation subset provides an additional 1,891 test samples. The benchmark covers seven major document VQA domains.

Source	Document Type	Train	Evaluation		
		# Q–D Pairs	# Q (% Preserved)	# D	# Pos. D per Q
Synthetic Degradation Dataset (12 degradations)					
ArXivQA [6]	Arxiv Figures	25,856	816 (8%)	8,066	1.00
ChartQA [7]	Charts	4,224	63 (5%)	500	1.00
MP-DocVQA [17]	Industrial Documents	10,624	591 (11%)	741	1.00
InfoVQA [8]	Infographics	17,664	718 (26%)	459	1.00
PlotQA [9]	Scientific Plots	56,192	863 (4%)	9,593	1.00
SlideVQA [16]	Slide Decks	8,192	556 (25%)	1,284	1.26
Synthetic [20]	Various	239,358	-	-	-
Real Degradation Dataset (5 degradations)					
ArXivQA [6]	Arxiv Figures	–	300 (10%)	3,000	1.00
MP-DocVQA [17]	Industrial Documents	–	591 (11%)	741	1.00
RVL-CDIP [3]	Various	–	1,000 (7%)	3,000	1.00

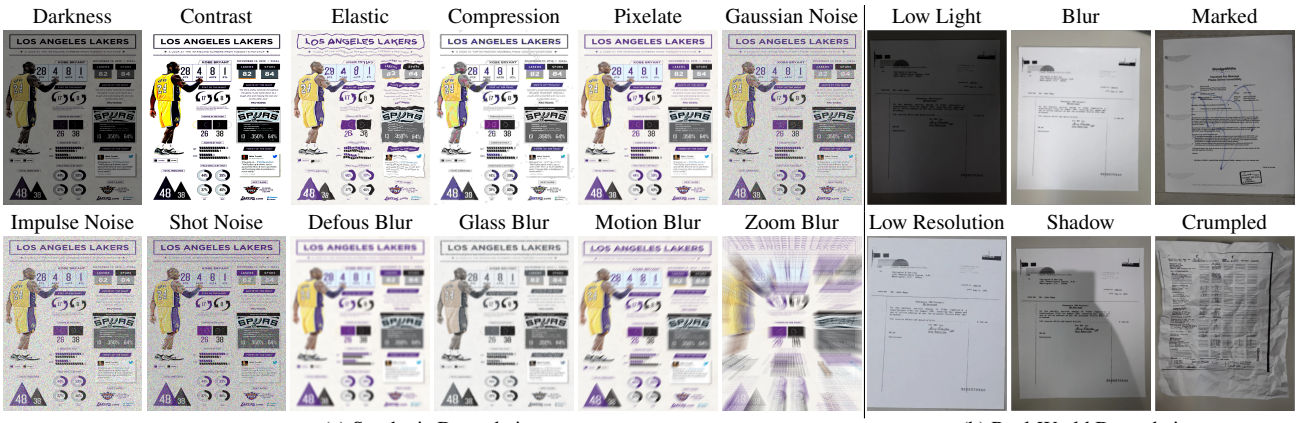
1.2. Real-world Degradation Pipeline

To evaluate robustness under realistic acquisition conditions, we construct a Real-World Degradation Dataset by physically printing and re-capturing document pages sampled from ArXivQA [6], MP-DocVQA [17], and RVL-CDIP [3]. We consider five types of real-world document degradations in our evaluation. All documents were recaptured using a Sony RX100 VII under controlled yet realistic conditions:

Blur. The camera shutter speed was reduced to $1/3$s and slight hand-induced motion was applied during exposure, producing realistic motion blur representative of handheld or unstable capture conditions.

Low light. The exposure compensation was set to -3EV to reduce incident illumination, resulting in low-light and low-contrast images typical of indoor or shaded environments.

Low resolution. Images were captured at a 24 mm focal length and subsequently cropped to reduce effective spatial resolution. This degrades character sharpness and introduces jagged text boundaries, mimicking long-distance or compressed document captures.



(a) Synthetic Degradation (b) Real-World Degradation

Figure 1. **Illustrative samples of images with synthetic and real degradations from the DVisRAG.**

Shadow. Localized shadows were introduced by partially blocking the light source (e.g., with a hand or objects), creating non-uniform illumination patterns similar to those encountered in everyday office or indoor settings.

Deliberate damage. Documents were physically altered by adding handwritten markings or by folding and crumpling the paper. These operations cause partial text occlusion and background irregularities, simulating user-altered documents and real-world wear.

To illustrate the visual characteristics introduced by these real-world distortions, Figure 1 presents example images for all five categories. These samples show that real degradations often contain complex and spatially varying effects, such as uneven illumination, non-rigid deformation, or mixed blur patterns, which more closely reflect the challenging conditions encountered in practical document understanding scenarios than synthetic perturbations.

1.3. OOD Benchmark Construction (RVL-CDIP)

To evaluate out-of-distribution generalization, we construct a zero-shot testing set based on the RVL-CDIP collection [3], a large-scale archive of official documents spanning 16 heterogeneous categories. Since RVL-CDIP is not a VQA dataset and contains no question–answer annotations, we generate new QA pairs specifically for this benchmark. We sample 200 pages per category (3,000 pages in total) and re-capture each page under the same five real-world imaging conditions described in Sec. 1.2. For each clean page image, we use GPT-4o to generate 3–5 factual, single-page VQA-style QA pairs using a constrained prompt that enforces document-grounded answers without external knowledge. We retain only items with a confidence score of at least 0.90 and further perform strict manual verification to ensure correctness, clarity, and answerability, removing any ambiguous or context-dependent cases. A stratified sampling procedure yields a final set of 1,000 high-quality QA pairs. These QA pairs are reused across all re-captured versions of the same page, enabling consistent evaluation of model robustness under diverse real-world degradations.

2. Implementation Details

RobustVisRAG is built upon MiniCPM-V 2.0 [11] as the retriever and MiniCPM-V 2.6 [12] as the generator. All experiments are conducted using PyTorch on eight NVIDIA A100 GPUs. The full RobustVisRAG pipeline consists of two training stages.

Stage 1: Retrieval. We initialize the retriever using the pre-trained weights from VisRAG [20]. To enable distortion-invariant semantic learning, we introduce a non-causal path into the vision encoder while adding a single non-causal token. During training, the temperature of the contrastive loss is set to 0.02, and the loss coefficients $(\lambda_{\text{FASL}}, \lambda_1, \lambda_2)$ are set to $(0.2, 0.5, 0.05)$. In addition, we apply the masking strategy in Eq. (9) & (12) at all Transformer layers and across all attention heads. We train the retriever for 2 epochs with AdamW, a learning rate of 1×10^{-5}, a batch size of 256, and a WarmupDecayLR scheduler. The total training timme for this stage is approximately 48 hours. During inference, the non-causal path is simply removed, and the encoder produces the semantic embedding Z_{sem} for retrieval. This ensures that the inference-time architecture and computational cost remain fully compatible with standard VisRAG-ret.

Stage 2: Generation. We adopt MiniCPM-V 2.6 [12] with pre-trained weights and keep the LLM backbone frozen. To enhance robustness against degraded visual inputs while preserving the zero-shot performance, we fine-tune only the vision encoder using an unsupervised training

```
# QUERY GENERATION PROMPT
# ========================================================================
# TASK: Generate 3-5 factual VQA questions from a single-page document image.

QUESTION TYPES:
- Dates, times, or time periods
- Names of people, organizations, or places
- Numerical values (percentages, amounts, quantities)
- Items in lists or tables
- Actions, purposes, or conditions stated in text

REQUIREMENTS:
1) Single-page constraint
   * Each question must be answerable from THIS page alone.
   * No cross-page or multi-document context.

2) Document-grounded only
   * Use ONLY information visible on the page.
   * Do NOT rely on world knowledge.
   * Answers must be directly verifiable from the document.

3) Answer format
   * Short factual phrases (1 to 10 words).
   * Preserve formatting for numbers/dates.

4) Question clarity
   * Direct, unambiguous wording.
   * Include a variety of question types (who / what / when / where / how much).

5) Confidence scoring
   * Evaluate legibility of the answer region.
   * Consider blur, skew, low contrast, etc.
   * Minimum confidence = 0.90.

OUTPUT (JSON only):
{
  "image_id": "filename",
  "questions": [
    {
      "qid": "unique_id",
      "query": "...",
      "answer": "...",
      "docid": ["document_identifier"],
      "prompt": "Answer the question using a single word or phrase.\nQuestion: ...\nAnswer:",
      "confidence": 0.90
    }
  ]
}
```

Figure 2. **Full prompt used for generating single-page VQA questions for OOD dataset.**

paradigm. This stage employs the Adam optimizer with cosine annealing, a learning rate of 2×10^{-6}, and a batch size of 256, trained for 3 epochs. The loss weight λ_3 is set to 0.03. This training stage requires approximaltely 26 hours. At inference time, the fine-tuned vision encoder replaces the original encoder of MiniCPM-V 2.6 [12], and the remaining generation process follows the standard VisRAG-gen pipeline without modification. We present the prompts for generation in Table 2.

3. Experimental Detail

3.1. Retrieval Performance

Table 3 provides the detailed retrieval results across all datasets. RobustVisRAG achieves consistent improvements over all baselines on both clean and degraded conditions. Notably, RobustVisRAG also achieves a 5.46% improvement on the out-of-domain RVL-CDIP benchmark, confirming the strong cross-domain generalization of our training strategy. Furthermore, RobustVisRAG surpasses the adversarially trained VisRAG-FARE [15] across all dataset, demonstrating that causal semantic–degradation disentan-

```
# QUERY FILTER PROMPT
# ========================================================================
# TASK: Validate VQA question and answer pairs for document images.

VALIDATION CHECKS:

1) Single-page constraint
    PASS: Answerable from this page alone.
    FAIL: Requires other pages or multi-hop context.
2) Document-grounded
    PASS: Answer is visibly present in the document.
    FAIL: Requires external knowledge or inference.
3) Answer verifiability
    PASS: The answer text clearly appears on the page.
    FAIL: Not directly visible or too ambiguous.
4) Question clarity
    PASS: Single clear interpretation.
    FAIL: Ambiguous or multiple plausible answers.
5) Confidence assessment
   * Evaluate legibility under degradation (blur, fade, skew).
   * Adjust confidence if needed.
   * Final confidence must be >= 0.90.

FAILURE REASONS:
- "cross_page"
- "external_knowledge"
- "not_verifiable"
- "ambiguous"
- "low_confidence"

OUTPUT (JSON only):
{
  "pass": true/false,
  "reason": "failure_reason or null",
  "adjusted_confidence": 0.00,
  "recommendation": "optional improvements"
}
```

Figure 3. **Prompt used to filter and validate the generated questions to ensure single-page answerability and unambiguous grounding.**

Table 2. **Prompt templates for generation. "Others" refers to all VQA datasets except ArxivQA.**

	ArxivQA	Others
VisRAG	{{ document(s) }} Question: {query }} Options: A. {{ Option 1 }} B. {{ Option 2 }} C. {{ Option 3 }} D. {{ Option 4 }} Answer directly with the letter of the correct option as the first character.	{{ document(s) }} Answer the question using a single word or phrase. Question:{{ query }} Answer:

glement is fundamentally more effective than perturbation-based robustness methods. These results highlight the strong generalization ability of the CSA module and its effectiveness in stabilizing retrieval performance under diverse and challenging real-world distortions.

3.2. Generation Performance

Table 4 presents the complete generation results across all synthetic and real-world datasets under different retrieval settings. RobustVisRAG consistently achieves the best performance across all conditions, such as on the Oracle setting, where it improves over VisRAG-Gen [20] by 6.35% and surpasses GPT-4o [10] by 10.42%. Notably, full finetuning (FFT) of VisRAG-Gen often leads to degraded performance on the out-of-domain RVL-CDIP benchmark, suggesting that supervised adaptation on VisRAG's distribution may introduce domain overfitting. In contrast, our method relies on an unsupervised degradation-guided objective that explicitly disentangles semantic and distortion factors without relying on dataset-specific annotations, enabling substantially stronger cross-domain generalization. These results confirm that the CSA module provides robust, degradation-invariant generation even under challenging synthetic distortions and real-world recaptures.

Table 3. **Overall retrieval performance in MRR@10** across clear, synthetic, and real-world degradation settings over nine benchmark datasets. For clarity, we note that Table 1 in the main text reports our retrieval model simply as RobustVisRAG. In this supplementary material, we use the notation RobustVisRAG-Ret to explicitly denote the retriever component; however, RobustVisRAG-Ret and the RobustVisRAG reported in Table 1 of the main text refer to the exact same model trained with our causality-guided framework. The best value in each column is in **bold**, and the second best is underlined.

Model (MRR@10)	# Para.	Synthetic Dataset														Real-World Dataset			
		ArxivQA		ChartQA		DocVQA		InfoVQA		PlotQA		SlideVQA		Average		RVL-CDIP	DocVQA	ArxivQA	Average
		clean	degra	clean	degra	clean	degra	clean	degra	clean	degra	clean	degra	clean	degra	degra	degra	degra	degra
BM25 (T) [14]	n.a	34.77	25.60	52.75	28.00	70.88	44.96	63.29	41.00	37.76	15.75	60.59	41.48	53.34	32.80	30.51	56.93	28.36	38.60
BGE-large (T) [18]	335M	36.76	25.85	55.43	29.72	56.09	39.65	76.50	55.60	47.33	19.45	81.75	52.02	58.98	37.05	26.00	48.52	31.19	35.23
NV-Embed-v2 (T) [5]	7.85B	53.11	39.29	73.66	41.83	73.92	52.49	84.06	65.20	57.36	24.71	92.38	62.56	72.41	47.68	37.52	65.10	45.71	49.44
MiniCPM-FV (T) [19]	2.72B	63.16	47.00	68.75	38.88	**79.20**	55.22	84.61	62.28	**65.46**	27.89	88.49	59.96	74.94	48.54	40.35	68.10	47.82	52.09
MiniCPM-FM (T) [19]	2.72B	62.40	47.41	68.53	43.84	75.83	54.64	81.97	62.68	63.45	30.28	86.83	60.89	73.17	49.96	39.14	66.90	49.11	51.72
SigLIP [21]	883M	22.82	20.19	59.27	38.25	34.33	29.10	54.60	43.33	26.86	15.16	62.91	52.57	43.47	33.10	5.02	20.49	19.71	15.07
SigLIP-FV [21]	883M	59.44	49.22	77.34	60.76	66.78	58.67	76.77	63.65	60.10	37.52	88.66	75.24	71.52	57.51	7.56	30.97	46.95	28.50
SigLIP-FM [21]	883M	60.23	54.64	78.19	64.69	65.70	59.57	74.32	64.67	48.38	35.20	86.24	79.84	68.84	59.77	9.39	36.35	49.93	31.89
ColPali-FV [2]	2.97B	73.87	63.94	72.70	58.95	78.55	70.31	79.37	64.79	51.84	28.80	91.10	80.53	74.57	61.22	15.73	48.47	60.91	41.70
VisRAG-Ret [20]	3.43B	74.58	64.63	75.65	61.99	75.21	68.99	86.18	77.62	61.75	40.29	92.04	82.22	77.57	65.96	41.74	65.42	62.27	56.48
VisRAG-Ret-FM [20]	3.43B	75.60	67.47	77.24	63.57	76.08	71.92	86.28	81.25	62.30	43.66	92.86	84.26	78.39	68.69	43.64	66.99	64.11	58.25
VisRAG-Ret-FM (FARE) [15]	3.43B	75.29	67.89	77.73	64.34	76.75	72.12	86.37	81.98	62.23	44.18	92.17	84.12	78.42	69.11	44.70	67.19	66.28	59.39
RobustVisRAG-Ret	3.43B	**78.75**	**76.30**	**80.01**	**69.95**	77.47	**74.83**	**87.74**	**83.26**	63.78	**48.80**	**92.93**	**86.11**	**80.11**	**73.21**	**47.20**	**72.38**	**71.89**	**63.82**

Table 4. **Overall generation performance in accuracy** across clear, synthetic, and real-world degradation settings over nine benchmark datasets. For clarity, we note that Table 2 in the main text reports our generation model simply as RobustVisRAG. In this supplementary material, we use the notation RobustVisRAG-Gen to explicitly denote the generator component; however, RobustVisRAG-Gen and the RobustVisRAG reported in the main text refer to the exact same model trained with our proposed causality-guided framework.

Model / Method	Input	Synthetic Dataset														Real-World Dataset			
		ArxivQA		ChartQA		DocVQA		InfoVQA		PlotQA		SlideVQA		Average		RVL-CDIP	DocVQA	ArxivQA	Average
		clean	degra	clean	degra	clean	degra	clean	degra	clean	degra	clean	degra	clean	degra	degra	degra	degra	degra
GPT-4o (T) [10]	top-1	59.74	57.48	46.98	12.70	50.16	22.50	45.18	33.57	14.45	3.82	47.77	27.52	44.05	26.26	5.20	19.29	58.33	27.61
	top-2	60.86	58.09	48.16	15.87	57.93	26.23	48.61	40.25	17.53	4.06	51.55	32.55	47.44	29.51	5.70	22.17	58.33	28.73
	top-3	62.50	57.23	45.16	19.05	55.81	28.43	50.86	40.95	14.88	5.56	52.99	32.55	47.03	30.63	6.80	23.52	58.00	29.44
	Oracle	61.89	60.91	66.67	26.98	63.79	46.87	50.56	46.10	20.39	10.54	55.04	35.97	53.06	37.90	24.70	39.93	60.00	41.54
MiniCPM (T) [19]	top-1	45.52	40.63	26.81	17.46	31.15	12.18	20.36	14.62	16.82	11.47	30.00	15.11	28.44	18.58	2.80	11.17	40.67	18.21
	top-2	42.46	41.54	24.05	15.87	35.55	14.21	20.36	7.66	17.17	11.82	31.00	17.27	28.43	18.06	3.70	12.52	44.00	20.07
	top-3	44.95	40.69	20.63	14.29	32.87	15.06	19.85	6.82	17.05	12.05	31.36	20.14	27.79	18.18	2.70	13.71	40.00	18.80
	Oracle	45.10	44.98	34.92	23.81	41.46	25.89	21.73	19.22	17.38	12.05	30.94	19.42	31.92	24.23	12.20	19.80	43.67	25.22
GPT-4o [10]	top-1	64.71	63.24	52.38	34.92	58.88	52.31	**63.09**	**53.99**	20.74	16.63	**54.86**	**40.81**	52.44	43.31	15.90	48.22	60.33	41.49
	top-2	63.36	62.50	49.21	39.68	64.13	55.73	**66.85**	55.62	20.16	13.44	**58.45**	44.78	53.69	45.29	17.10	54.15	61.00	44.08
	top-3	62.01	63.48	53.97	38.10	67.17	55.99	**66.43**	56.10	19.35	13.09	**60.97**	47.66	54.98	45.74	17.80	53.64	63.00	44.81
	Oracle	66.05	70.66	68.25	**53.14**	79.36	70.47	**71.45**	60.30	31.29	19.33	**64.57**	53.23	63.50	54.52	37.10	71.07	67.67	58.61
VisRAG-Gen [20]	top-1	**68.63**	67.77	52.38	33.33	63.28	52.96	53.06	41.23	23.87	20.05	47.84	37.61	51.51	42.16	19.20	48.90	66.00	44.70
	top-2	69.00	66.05	57.14	34.92	69.37	58.04	52.65	41.92	24.91	19.93	49.82	39.21	53.82	43.35	21.80	54.99	64.67	47.15
	top-3	68.14	65.32	57.14	38.10	72.76	61.42	53.20	42.76	23.99	20.74	49.46	38.85	54.11	44.53	23.60	55.16	62.67	47.14
	Oracle	71.69	70.81	68.25	42.86	83.08	69.71	62.67	49.44	36.73	28.16	57.73	47.12	63.36	51.52	47.00	70.05	71.00	62.68
VisRAG-Gen-FM (PEFT) [4]	top-1	68.19	68.10	53.97	35.27	71.01	58.45	56.97	42.68	23.81	20.14	49.57	37.18	53.92	43.64	21.17	53.81	66.67	47.22
	top-2	69.19	67.10	57.19	34.34	74.19	61.08	54.80	43.68	**27.17**	20.08	50.10	40.78	55.44	44.51	23.94	55.18	67.33	48.82
	top-3	67.97	66.10	59.89	37.01	75.81	63.78	53.28	43.17	24.28	21.20	50.10	39.81	55.22	45.18	27.19	58.19	62.67	49.35
	Oracle	71.98	71.27	70.19	43.19	84.19	70.74	64.79	52.18	38.97	28.40	59.18	49.91	64.88	53.16	47.18	72.18	71.20	63.52
VisRAG-Gen-FM (FFT) [20]	top-1	67.65	67.40	57.14	33.33	74.96	61.59	57.24	43.87	24.57	20.86	51.44	39.57	55.50	44.44	16.10	58.45	67.33	47.29
	top-2	68.63	67.28	58.73	34.92	78.17	64.47	56.55	45.54	24.57	20.51	51.26	42.81	56.32	45.92	20.60	59.18	**68.00**	49.59
	top-3	67.89	66.42	55.56	36.51	80.54	66.67	54.31	44.56	24.68	22.25	51.08	40.64	55.68	46.17	21.60	61.81	63.67	49.03
	Oracle	72.79	70.19	71.89	44.19	85.19	72.01	68.18	53.81	37.19	28.84	60.19	50.18	65.91	53.20	42.37	73.18	72.33	62.54
VisRAG-Gen-FM (FARE) [15]	top-1	68.10	68.10	59.18	34.18	73.13	62.10	58.24	46.60	24.18	20.42	50.10	39.68	55.49	45.68	25.10	61.76	66.00	50.95
	top-2	68.20	68.43	60.42	36.26	77.18	65.43	57.36	48.72	26.80	21.53	51.26	46.10	56.87	47.75	28.60	65.14	65.67	53.14
	top-3	68.10	67.19	61.06	37.53	79.70	66.82	56.85	49.52	25.60	20.21	54.86	47.80	57.02	48.18	29.60	63.45	67.33	53.46
	Oracle	72.19	71.20	71.96	48.98	85.15	72.18	68.10	54.19	38.10	29.18	61.28	52.42	66.13	54.69	47.20	75.17	71.00	64.46
RobustVisRAG-Gen	top-1	68.32	**69.07**	**63.49**	**39.68**	**76.99**	**65.99**	60.19	52.23	**27.11**	**21.42**	53.24	39.75	**58.22**	**48.02**	**27.10**	**70.39**	**68.67**	**55.39**
	top-2	68.18	**68.95**	**66.67**	**49.21**	**85.11**	**71.40**	63.68	**57.52**	26.77	**22.02**	55.49	**50.00**	**60.98**	**53.18**	**29.40**	**76.65**	67.67	**57.91**
	top-3	**68.63**	**69.56**	**67.25**	**47.62**	**87.14**	**72.93**	65.77	**59.05**	**27.11**	**24.71**	56.01	**50.18**	**61.99**	**54.01**	**30.30**	**79.70**	**68.33**	**59.44**
	Oracle	72.49	**71.28**	**72.60**	52.38	**85.79**	**74.79**	67.72	**60.31**	**41.60**	**31.32**	63.80	**57.12**	**67.33**	**57.87**	**50.80**	**81.29**	**75.00**	**69.03**

Table 5. **Overall end-to-end performance** under clean settings across six benchmark datasets.

Methods	Retrieval (MRR10)							Generation (Top-1)						
	ArxivQA	ChartQA	DocVQA	InfoVQA	PlotQA	SlideVQA	Average	ArxivQA	ChartQA	DocVQA	InfoVQA	PlotQA	SlideVQA	Average
VisRAG	74.58	75.65	75.21	86.18	61.75	92.04	77.57	66.30	46.03	61.25	55.15	24.22	49.46	50.40
VisRAG-FT	75.29	77.73	76.75	86.37	62.23	92.17	78.42	67.50	57.00	72.76	57.36	24.57	49.82	54.84
Two-stage	74.78	76.50	75.29	86.42	61.76	91.93	77.78	66.54	47.62	61.76	55.43	23.99	48.02	50.56
RobustVisRAG	**78.75**	**80.01**	**77.47**	**87.74**	**63.78**	**92.93**	**80.11**	**68.32**	**63.49**	**76.99**	**60.19**	**27.11**	**53.24**	**58.22**

Table 6. **Overall end-to-end performance** under synthetic degradation settings over six benchmark datasets.

Methods	Retrieval (MRR10)							Generation (Top-1)						
	ArxivQA	ChartQA	DocVQA	InfoVQA	PlotQA	SlideVQA	Average	ArxivQA	ChartQA	DocVQA	InfoVQA	PlotQA	SlideVQA	Average
VisRAG	64.63	61.99	68.99	77.62	40.29	82.22	65.96	66.42	33.33	51.61	42.76	20.74	36.87	41.96
VisRAG-FT	67.89	64.34	72.12	81.98	44.18	84.12	69.11	67.77	35.98	61.08	44.56	20.86	39.48	44.96
Two-stage	65.53	62.11	69.91	78.23	41.34	81.82	66.49	66.79	34.92	51.10	43.18	20.28	37.23	42.25
RobustVisRAG	**76.30**	**69.95**	**74.83**	**83.26**	**48.80**	**86.11**	**73.21**	**69.07**	**39.68**	**65.99**	**52.23**	**21.42**	**39.75**	**48.02**

Table 7. **Overall end-to-end performance** under real-world degradation settings over three benchmark datasets.

Methods	Retrieval (MRR10)				Generation (Top-1)			
	RVL-CDIP	DocVQA	ArxivQA	Average	RVL-CDIP	DocVQA	ArxivQA	Average
VisRAG	41.74	65.42	62.27	56.47	18.90	48.39	61.67	42.99
VisRAG-FT	44.70	67.19	66.28	59.39	21.70	57.10	66.00	48.27
Two-stage	34.86	64.87	61.05	53.59	17.00	39.26	65.00	40.42
RobustVisRAG	**47.20**	**72.38**	**71.89**	**63.82**	**27.10**	**70.39**	**68.67**	**55.39**

Table 8. **Analysis of the weighting coefficients for retrieval.**

$(\lambda_{\text{FSAL}}, \lambda_1, \lambda_2)$	Adjust CSA			Balance	Adjust NCDM		
	(0.0, 0.5, 0.05)	(0.2, 0.05, 0.05)	(0.2, 0.7, 0.05)	(0.2, 0.5, 0.05)	(0.2, 0.5, 0.5)	(0.2, 0.5, 0.07)	(0.2, 0.5, 0.005)
Synthetic	71.89	70.19	72.98	**73.21**	68.18	73.18	69.76
Real-world	61.92	60.24	63.09	**63.82**	58.89	63.17	59.17

Table 9. **Analysis of the weighting coefficients for generation.**

λ_3	0.3	0.03	0.003
Synthetic	45.32	**48.02**	46.67
Real-world	51.79	**55.39**	52.23

Table 10. **Analysis of the number of non-causal tokens.**

Number of Z_{deg}	1	2	4
Retrieval (MRR@10)	**63.82**	63.74	62.93
Generation (Top-1)	**55.39**	55.01	54.32

Table 11. **Analysis of Degradation Representation.**

Strategy	w/o NCDM	w/ NCDM
mAP	24.9	92.3

Table 12. **Analysis of Additional Two-Stage Pipelines.**

Configurations	Retrieval (MRR@10)			Generation (Top-1)		
	VisRAG	Synthetic	Real	VisRAG	Synthetic	Real
VisRAG	77.57	65.96	56.47	50.40	41.96	42.99
PromptIR+VisRAG	77.78	66.49	53.59	50.56	42.25	40.42
UniRestore+VisRAG	78.21	68.28	58.27	51.72	44.68	46.26
RobustVisRAG	**80.11**	**73.21**	**63.82**	**58.22**	**48.02**	**55.39**

3.3. End-to-End Performance

Tables 5–7 present the full end-to-end evaluations under clean, synthetic, and real-world degradation settings. RobustVisRAG achieves consistent improvements across all conditions, surpassing VisRAG and VisRAG-FT on clean datasets while demonstrating substantial gains under degraded inputs. We also observe that the Two-stage strategy [13], which restoring images before feeding them into VisRAG, does not guarantee improvement, as the restoration step may introduce artifacts or distort clean images. Under real-world degradations, RobustVisRAG improves retrieval accuracy by 7.35% on average and further boosts end-to-end generation accuracy by 12.40% compared with VisRAG. Notably, transferring from synthetic training to real-world degradations on DocVQA (e.g., 65.99 $\rightarrow$ 70.39), confirming that the benefits of semantic–degradation disentanglement propagate throughout the entire pipeline. These results highlight the practical robustness of RobustVisRAG and its ability to maintain reliable multimodal reasoning in realistic imaging environments.

3.4. Investigation of Hyperparameter

Table 8 and 9 summarize the effects of different loss-weight settings. We observe that removing the FASL term ($\lambda_{\text{FASL}} = 0$) reduces performance, indicating that the fine-grained alignment term provides important detail supervision beyond the SIL loss. λ_1 and λ_2 either overemphasize degradation modeling or weaken the non-causal pathway, both leading to suboptimal results. Across a reasonable range of weighting coefficients, performance varies smoothly and remains stable under degradations. A balanced configuration of (0.2, 0.5, 0.05) achieves the best retrieval performance. A similar trend is found on the generation side, where overly large or small λ_3 reduce its robustness. We set $\lambda_3 = 0.03$ as the optimal trade-off. Notably, although training is more complex than standard fine-tuning, our method introduces no additional inference cost and consistently improves robustness.

3.5. Number of Non-causal Tokens

Table 10 shows that increasing the number of Z_{deg} does not consistently improve performance on the real DVisRAG subset, as multiple non-causal tokens capture overlapping degradation cues, leading to redundant rather than more informative representations.

3.6. Analysis of Degradation Representation

We conduct an additional experiment to quantitatively evaluate the structure induced by NCDM in the Z_{deg} space. We synthesize four degradation types (darkness, blur, low resolution, and shadow) on 100 clean slide images from SlideVQA as queries, and probe the real-degradation subset in DVisRAG. For each query, we retrieve the top-5 images with the smallest distance in the Z_{deg} embedding space and compute mean average precision (mAP) based on degradation-type labels. Table 11 shows that incorporating NCDM improves mAP, indicating that Z_{deg} captures degradation-consistent and quantitatively discriminative structure that supports task-level retrieval of images with similar degradation attributes.

3.7. Analysis of Additional Two-Stage Pipelines

We additionally include UniRestore [1]+VisRAG, a task-oriented restoration pipeline, for comparison. Table 12 shows that under our evaluated VisRAG setting, task-oriented restoration does not close the performance gap to RobustVisRAG under degradations.

3.8. Case Study

As shown in Figure 4, demonstrates that RobustVisRAG remains reliable even under real-world degradations, including low-light conditions, shadow occlusion, and physical document damage. While VisRAG often retrieves visually similar but semantically incorrect pages, our method

consistently identifies the correct ground-truth document by leveraging distortion-invariant semantic features and disentangled degradation cues. This demonstrates the strong robustness of RobustVisRAG in challenging retrieval scenarios and confirms its advantage in maintaining accurate evidence retrieval for downstream multimodal reasoning.

As shown in Figure 5, RobustVisRAG also demonstrates clear advantages on the generation side. Even when both models are provided with the same retrieved document, VisRAG often struggles to extract key information from heavily degraded pages, producing incomplete or incorrect answers. In contrast, RobustVisRAG can reliably parse text and semantic content from damaged, low-light, or shadowed documents, leading to accurate and faithful responses. These results highlight the importance of robust visual understanding during generation and further validate the effectiveness of our distortion-aware design in real-world document QA scenarios.

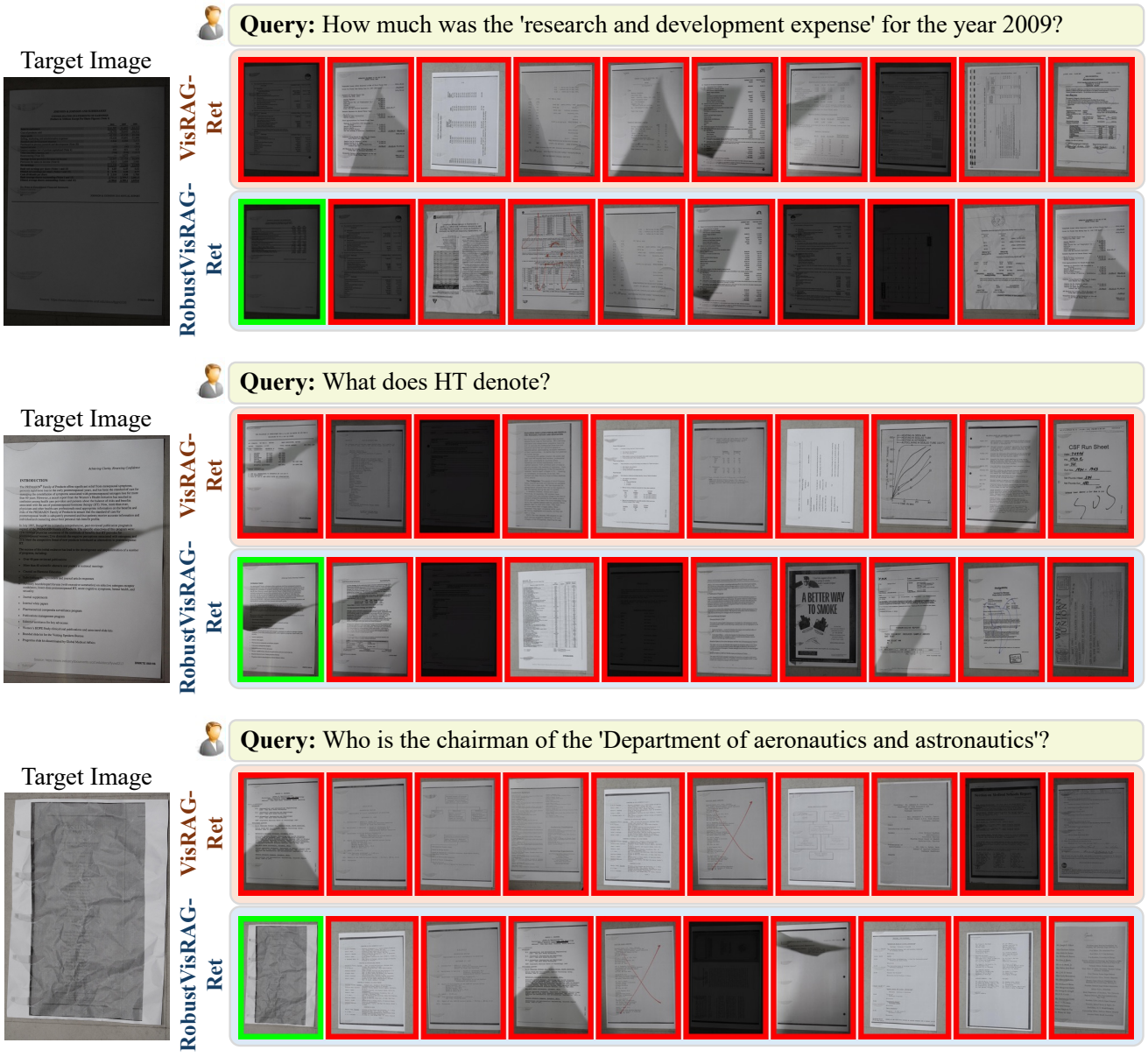


Figure 4. **Reteieval case studies from DVisRAG under low-light, shadow, and damaged degradation.** The results demonstrate VisRAG retrieves visually similar but semantically irrelevant documents, while RobustVisRAG consistently retrieves the correct ground-truth target despite severe distortions. These results demonstrate the robustness of the proposed method to illumination changes, shadows, and physical damage.

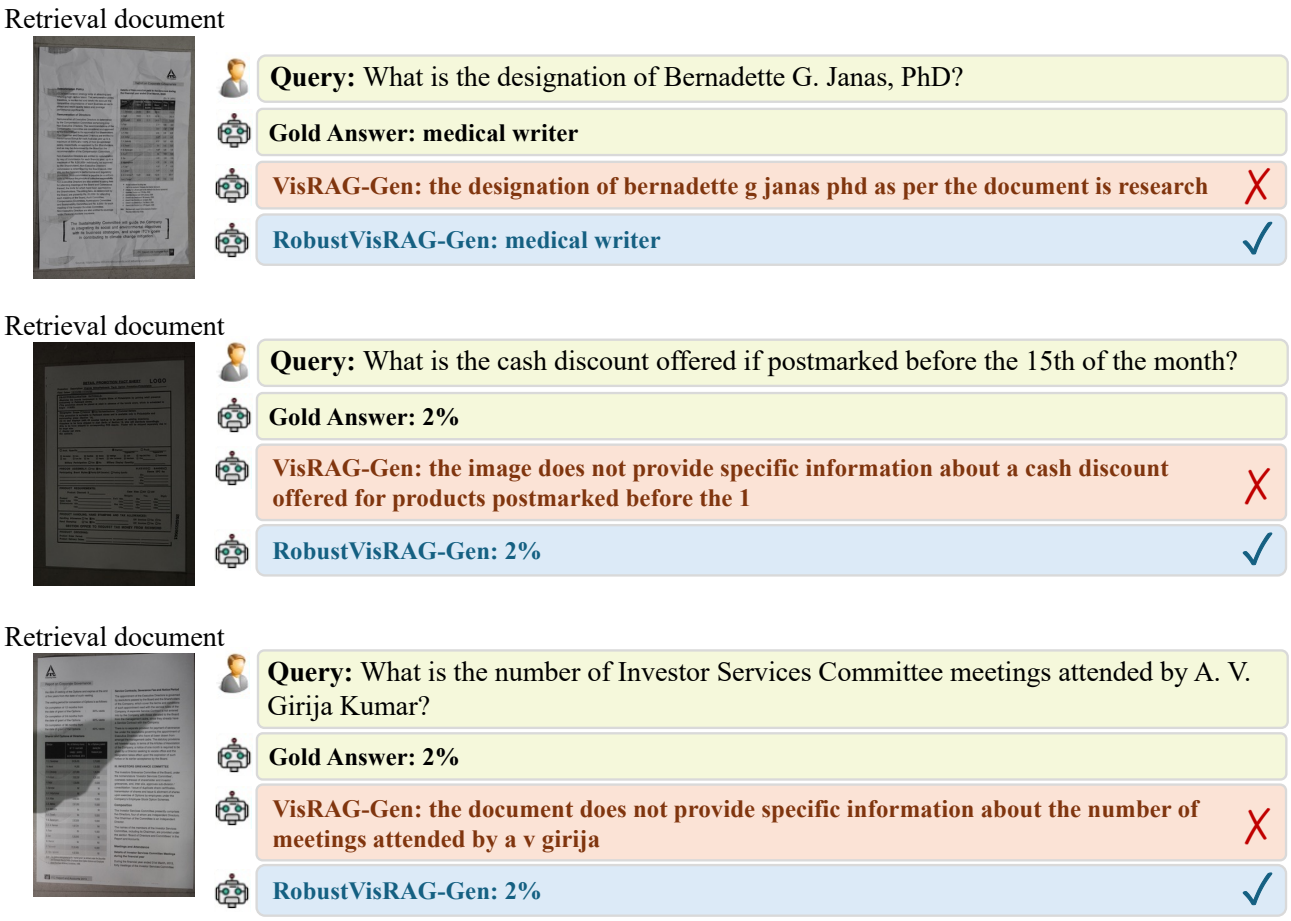


Figure 5. **Generation case studies from DVisRAG under damaged, low-light and shadow degradation.** Even if VisRAG is given the correct document during generation, it often fails to extract reliable information from heavily degraded inputs, leading to incorrect or incomplete responses. In contrast, RobustVisRAG can accurately interpret the degraded document and produce correct answers, demonstrating its resilience to real-world distortions.

References

[1] I Chen, Wei-Ting Chen, Yu-Wei Liu, Yuan-Chun Chiang, Sy-Yen Kuo, Ming-Hsuan Yang, et al. Unirestore: Unified perceptual and task-oriented image restoration model using diffusion prior. In *CVPR*, 2025. 1, 6

[2] Manuel Faysse, Hugues Sibille, Tony Wu, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*, 2024. 5

[3] Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval. In *ICDAR*, 2015. 1, 2

[4] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ICLR*, 2022. 5

[5] Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*, 2024. 5

[6] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *ACL*, 2024. 1

[7] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *ACL*, 2022. 1

[8] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *WACV*, 2022. 1

[9] Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *WACV*, 2020. 1

[10] OpenAI. Hello, gpt-4o, 2024. 4, 5

[11] OpenBMB. openbmb/minicpm-v-2, 2024. 2

[12] OpenBMB. openbmb/minicpm-v-2_6, 2024. 2, 3

[13] Vaishnav Potlapalli, Syed Waqas Zamir, Salman Khan, and Fahad Khan. Promptir: Prompting for all-in-one image restoration. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 6

[14] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. Okapi at trec-3. In *TREC-3*, 1995. 5

[15] Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. *ICML*, 2024. 3, 5

[16] Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. Slidevqa: A dataset for document visual question answering on multiple images. In *AAAI*, 2023. 1

[17] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 2023. 1

[18] Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*, 2023. 5

[19] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 5

[20] Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, et al. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *ICLR*, 2025. 1, 2, 4, 5

[21] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. 5