# FontHalu: Font-based Hallucinations in Multimodal Large Language Models

Anonymous ACL submission

#### Abstract

Multimodal large language models (MLLMs) 001 have achieved remarkable performance in processing and reasoning over text and images. However, they remain susceptible to hallucinations-instances where generated content deviates from input data or contradicts established knowledge. While hallucinations in MLLMs have attracted increasing attention, the specific impact of font variation-a common yet overlooked source of hallucination-has not been systematically investigated. Moreover, existing OCR benchmarks include limited font diversity and primarily focus on layout or background changes, lacking fine-grained control over font factors and neglecting long-tail fonts. To address this gap, we introduce and categorize font-induced hallucinations, and conduct 017 comprehensive experiments to examine how fonts affect MLLMs across dimensions such as font perturbations, style shifts, font-semantic interactions, and sentiment recognition. Based on these findings, we propose FontHalu, a benchmark with diverse font types and scenario settings, specifically designed to evaluate MLLMs' robustness in OCR, key information extraction (KIE), and sentiment analysis under font variation. We will release FontHalu and related code to support research on improving the reliability and robustness of MLLMs.

#### 1 Introduction

037

041

MLLMs excel in text and image processing but suffer from hallucinations, producing inaccurate or unfounded content that compromises reliability. Existing studies have identified various forms of hallucinations in MLLMs. For instance, Liu et al. (2024d) demonstrate that MLLMs can generate incorrect answers despite correctly interpreting visual content, using paired positive and negative visual question-answer samples. Similarly, PhD (Liu et al., 2024c) highlights significant variability in MLLM performance across different tasks, expos-



Figure 1: Examples of font-related hallucinations in MLLMs: a case study of Qwen2-VL-7B (Q) and MiniCPM-V-2.6 (M). The letters marked in red indicate those that are inconsistent with the ground truth.

ing inconsistencies in how these models reason over multimodal inputs.

While extensive research has explored general hallucinations in MLLMs, the interaction between textual and visual features introduces additional challenges. In particular, font styles and sizes can significantly affect MLLMs' perception of textual information. As shown in Figure 1, changes in font style may cause MLLMs to confuse visually similar letters during text recognition. Furthermore, stylistic variations can alter the perceived sentiment of a text, resulting in misclassification in sentiment analysis. We define this phenomenon as FontHalu—errors or unreasonable outputs generated by MLLMs due to the visual characteristics of fonts within input images, rather than the textual semantics alone. Notably, FontHalu is not a dis-

0

(

077

0

0

089 090 091

> 092 093

09

09

099 100

101 102

102

104

105 106

107

108 109 tinct category separate from typical hallucination types such as category, attribute, and relationship. Instead, it serves as a complementary and extended subset within these categories, specifically addressing font-related issues.

FontHalu is not merely an academic concern but pose tangible risks in real-world applications. For example, in medical diagnostics, font misinterpretations in patient records may result in incorrect treatments, while in legal document processing, hallucinations could lead to misread contract clauses with significant consequences. Although the increasing reliance on MLLMs for automated document understanding, the effects of font variations remain largely unexplored, leaving a critical gap in model robustness evaluation.

Despite the significance of this issue, existing benchmarks fail to systematically assess the impact of font hallucinations on model reliability like CHAIR (Rohrbach et al., 2018) and POPE (Li et al., 2023c). Datasets such as OCR-Bench (Liu et al., 2024e), SROIE (Huang et al., 2021), DocVQA (Mathew et al., 2021), and TextVQA (Singh et al., 2019) have not systematically investigated the impact of font variations on OCR performance. Some benchmarks include recognition of handwritten or artistic fonts (Cui et al., 2024), they cover only a limited range of font variations and do not systematically investigate font-based hallucinations, leaving an important gap in MLLMs evaluation.

To address this gap, we first adopt a controlled variable methodology to systematically analyze MLLM performance across multiple dimensions of font variation, starting from basic font attributes such as size and weight, followed by overall style changes, more complex layout transformations, the semantic information carried by fonts, and finally their impact on sentiment recognition. These carefully designed experiments provide valuable insights into the challenges posed by font-based hallucinations. Based on these findings, we introduce the FontHalu benchmark-the first dedicated benchmark for evaluating MLLMs' robustness against font-based hallucinations. Finally, we also present a preliminary exploration of methods to mitigate font hallucinations. Our contributions are as follows:

1. We introduce the FontHalu phenomenon, categorize its types, and conduct a detailed empirical analysis to characterize its impact on

### MLLMs.

2. We develop FontHalu benchmark, the first benchmark specifically designed to systematically evaluate font-induced hallucinations in MLLMs, enabling standardized assessment and comparison.

110

111

112

113

114

115

116

117

118

119

120

121

122

123

124

125

126

127

128

129

130

131

132

133

134

135

136

137

138

139

140

141

142

143

144

145

146

147

148

149

150

151

152

153

154

155

156

157

158

3. We propose a font-clustering-based approach to construct a dedicated training dataset for MLLM fine-tuning, thereby mitigating font hallucinations in MLLMs.

## 2 Related Work

Hallucinations in MLLMs The remarkable success of LLMs has paved the way for the development of multimodal large language models (Li et al., 2023a; Liu et al., 2024b; Wang et al., 2024; Zhu et al., 2024), which combine pretrained visual models with LLMs to enable their visual capabilities. But at the same time, it also introduced many vision specific hallucination phenomena (Gunjal et al., 2024; Jiang et al., 2024; Zhou et al., 2024). These hallucinations can be roughly classified into category, attribute, and relationship (Bai et al., 2024). In the research on category types (Liu et al., 2024a; Yu et al., 2024) mainly focuses on object existence and descriptive issues, while research on fonts is relatively scarce. However, fonts play a crucial role in tasks such as KIE (Knowledge-Intensive Extraction) recognition, text-guided VQA (Visual Question Answering) tasks, and sentiment analysis tasks (Liu et al., 2023; Huang et al., 2021; Mathew et al., 2021; Singh et al., 2019). Therefore, studying the hallucinations problem caused by fonts is of great significance and can promote further development in the field of MLLM hallucinations.

**Benchmarks for MLLMs** There are many hallucination benchmarks for MLLMs, such as POPE (Li et al., 2023b), Nope (Lovenia et al., 2023), and CIEM (Hu et al., 2023). While effective for basic hallucination assessment, these benchmarks notably exclude OCR recognition capabilities from their evaluation scope (Liu et al., 2024a; Chen et al., 2023). Recently, more and more hallucination benchmarks have mentioned OCR recognition as an important part of multimodal hallucination assessment, such as Hallusion-Bench (Guan et al., 2024) and MME (Cui et al., 2024). However, their treatment of OCR-related hallucinations is limited, particularly in addressing the impact of font variations on hallucinations

- 159 160
- 161

166

167

169

170

171

172

173

174

175

176

177

in MLLMs, a factor that can significantly affect recognition accuracy.

#### **Font-based Hallucinations** 3

In this work, we identify three typical types of 162 FontHalu: hallucinations of font style, hallucinations of font semantics, and hallucinations of font 164 sentiment. 165

Hallucinations of Font Style From the perspective of a single font, Brown (2024) categorizes font variation axes into six dimensions: weight, width, slant, italic, optical size, and X-height. In this work, we focus on six key dimensions of font attributes: size, spacing, slant, alignment, weight, and X-height. In this work, we focus on six key dimensions of font attributes: size (from small to large), spacing (from tight to loose), slant (tilting letters to the right), alignment (left, right, or center), weight (from light to bold), and line-height (vertical space between lines).

The combination of dimensional variations re-178 sults in style changes within the same font. From 179 the perspective of different font families, font styles can vary widely, including categories such 181 as Serif (Arditi and Cho, 2005), Sans-Serif, Script, Monospaced, and others characterized by unique 183 features like strokes. These varying font styles affect MLLMs' ability to accurately recognize text, 185 leading to differences in performance across font types. We refer to this phenomenon as Font Style Hallucinations. For instance, transitioning from 188 189 a standard typeface to a more decorative or handwritten style can cause the model to misinterpret 190 the text, potentially leading to hallucinations and 191 incorrect inferences. In Figure 1, for example, the model confuses '7' with 'F' and 'g' with 'q' due to the specific font. This misinterpretation often mani-194 fests in the confusion of visually similar characters, 195 which is exacerbated by stylistic elements. 196

**Hallucinations of Font Semantics** In practical 197 applications, adjustments to font style can introduce additional semantic cues or nuances to the 199 text. For example, highlighting a word serves to emphasize its importance. Wu and Yuan (2003) find that highlighting can significantly improve table search efficiency. In handwritten documents, strike-through text may appear, potentially compli-204 cating OCR system recognition (Adak and Chaudhuri, 2014). This is particularly relevant since striking through a word and replacing it with a new 207

one signifies that the original term is invalid and has been superseded. Additionally, overlay text plays a vital role in video content analysis, providing key semantic cues for tasks such as video information retrieval and summarization (Adak and Chaudhuri, 2014). Similarly, when a word is written in a light color and overlaid with a darker one, the darker color often conveys the final or correct answer. However, MLLMs may not always be capable of recognizing these semantic cues, leading to hallucinations, which we refer to as Font Semantics Hallucinations. As shown in Figure 1, MiniCPM-V-2.6 (Yao et al., 2024) treats the strikethrough text as the final answer, resulting in a semantic hallucination.

208

209

210

211

212

213

214

215

216

217

218

219

220

222

223

224

225

226

227

228

229

230

231

232

233

234

235

236

237

238

239

240

241

242

243

244

245

246

247

248

249

250

251

252

253

254

255

256

257

**Hallucinations of Font Sentiment** Zhang et al. (2018) used semantic differential and statistical analysis methods to identify factors like exquisiteness, constriction, and a sense of order, which shape the emotional perception of fonts. Emotiondriven attention facilitation is influenced not only by biological relevance but also by perceptual features (Bayer et al., 2012). For example, fonts themselves can evoke emotional tones, such as the horror effect conveyed by the Silent Hill font. This raises the question: do MLLMs recognize these emotional cues in fonts? If so, they may misinterpret or exaggerate the emotional tone, resulting in what we term Font Sentiment Hallucinations. As shown in Figure 1, the models can correctly assess sentiment with a neutral (or standard) font. However, when we switch to a more cheerful-looking font, the models' response changes from negative to neutral. Similarly, when the models assess sentiment as positive in a neutral font, switching to a horror-themed font causes the sentiment evaluation to shift to neutral, resulting in a font sentiment hallucination.

We will experimentally validate and analyze the aforementioned hallucinations in Section 5.1.

#### 4 **FontHalu Benchmark**

#### 4.1 Overview

The FontHalu benchmark is a comprehensive evaluation framework designed to assess the impact of font-induced hallucinations on MLLM. This benchmark specifically targets the three types of font hallucinations introduced earlier: Font Style, Font Semantics, and Font Sentiment. To systematically evaluate these phenomena, we introduce three dedicated sub-tasks:



Figure 2: Benchmark Construction Pipeline. We first cluster and select 1,000 representative fonts from 19,247 candidates. Next, we construct 200 image templates and generate corresponding textual content. Finally, we build a benchmark dataset for three distinct tasks using the selected fonts, templates, and content.

- Font Variation Recognition (VarRec): This task involves extracting text rendered in different fonts from images with a blank background. It aims to measure the model's robustness to font variations in isolated text recognition scenarios.
- Font Semantic Analysis (SemAna): This task evaluates the model's ability to extract text from real-world contexts, such as shipping labels or documents, where additional semantic cues—such as strikethroughs, highlights, or overlays—may alter the intended meaning.
- Font Sentiment Analysis (SentAna): This task assesses the extent to which a model can correctly interpret the sentiment of a sentence when presented in a specific font, reflecting the potential influence of typography on sentiment perception.

#### 4.2 Construction

261

262

263

272

273

274

276

The benchmark construction process, illustrated in
Figure 2, consists of four key stages: font selection,
image template creation, content generation, and
benchmark assembly.

Font Selection In this stage, we collect a diverseset of fonts, including both commonly used and

rare artistic or commercial fonts, to ensure comprehensive coverage of font variations. Ultimately, we gather 19,274 distinct fonts, each applied to render identical content on  $250 \times 1500$  pixel blank images. The content consists of two lines of Lorem Ipsum text and one line of numerical data, generating a total of 19,274 images. These images are then processed using a pre-trained ViT model (ViT-Base-Patch16-224) (Dosovitskiy et al., 2021) to extract feature vectors from the final hidden layer. To reduce dimensionality, we apply UMAP (McInnes and Healy, 2018), followed by K-means clustering (Lloyd, 1982) to group the feature vectors into 200 clusters. From each cluster, five representative fonts are selected, yielding a total of 1,000 candidate fonts for the benchmark. For a detailed explanation of the font selections algorithm, see appendix A.

284

285

287

290

291

292

293

294

297

298

299

300

302

303

304

305

306

307

308

309

310

311

312

**Template Creation** In this phase, 200 images are selected from publicly available datasets, including DocVQA (Mathew et al., 2021), FUNSD (Jaume et al., 2019) and SROIE (Huang et al., 2021). We then annotate each image with 2–3 bounding boxes (annotation criteria are shown in Appendix B). To facilitate controlled content placement, we binarize these images by removing the content within the annotated regions while preserving the location information. This process results in 200 annotated template images.

**Content Generation** In this phase, we design various data types based on the annotated bounding boxes, including Lorem Ipsum text, numbers, dates, times, names, addresses, email addresses, etc. These content types are selected to reflect diverse real-world text data, ensuring broad scenario coverage. We utilized a data generation library to auto-generate diverse text contents.

313

314

315

319

322

324

325

327

331

337

338

339

340

341

342

343

344

345

347

349

For each task, we follow specific content generation strategies: **VarRec Task:** The selected fonts are used to write Lorem Ipsum content on blank images. **SemAna Task:** Questions are first generated based on the key information associated with the annotated bounding boxes in the image templates. Then, specific content is written in these bounding boxes using various fonts while applying one of four randomly selected formatting modes: strikethrough, highlighting, overlay, or normal. These modified texts serve as our groundtruth answers. **SentAna Task:** The test set of the Sentiment140<sup>1</sup> dataset is rendered in different fonts.

> Finally, we manually verify the dataset, removing low-quality samples to ensure data integrity (the specific criteria are shown in Appendix B). This results in the final version of the FontHalu benchmark. Statistics of the FontHalu benchmark are presented in Table 1.

Task	Images	Questions
VarRec	250	250
SemAna	780	1877
Normal	589	928
Strikethrough	124	132
Overlay	317	393
Highlight	343	424
SentAna	200	200

Table 1: Task overview: number of images and questions.

#### 4.3 Evaluation Metrics

We evaluate MLLMs' performance using the following metrics: ACC: Measures the presence of the expected answer in the generated response, with higher values indicating better performance. NED: Measures the normalized edit distance between the expected and generated answers, with lower values indicating better performance. Specific calculation formulas are in the appendix C.

#### **5** Experiments

#### 5.1 Font-based Hallucinations

To validate the three types of hallucinations introduced in Section 3, we design targeted experiments to analyze the characteristics and manifestations of FontHalu. 350

351

352

353

355

356

357

359

360

361

362

363

364

365

366

367

369

370

371

372

373

374

375

376

377

378

379

381

383

384

385

387

390

391

392

394

395

Settings When investigating hallucinations of font style, we first explore the variations in font attributes using a single, commonly used font (Roboto, one of the most popular fonts on Google Fonts). These attributes include size (ranging from 15 to 50 pt), spacing (from -0.1 to 0.4 em), slant (from 15 to 50 degrees), alignment (left, center, right), weight (from 100 to 900 on the font weight scale), and line height (from 0.4 to 2.0 em). To minimize potential confounding factors, we present the font against a plain background, avoiding any layout or contextual influences. This experiment is referred to as Font Style (Plain). Next, we examine the impact of different font types, using five distinct fonts. To simulate real-world scenarios more effectively, we incorporate these fonts into contextual settings (such as tracking numbers and receipts). This experiment, conducted on both Chinese and English datasets, is called Font Style (Scene).

For hallucinations of font semantics, We examine how MLLMs interpret content when text is presented in four different formatting styles: normal, overlay, highlight, and strikethrough. We refer to this experiment as **Font Semantics**.

We further examine how MLLMs exhibit **hallucinations of font sentiment** when presented with identical content rendered in five different fonts: a visually neutral font (NF), two horror-themed fonts (HF-A, HF-B), and two visually cheerful fonts (CF-A, CF-B)<sup>2</sup>. We refer to this experiment as **Font Sentiment**.

For detailed experimental settings (including steps, examples & prompts), refer to Appendix D. The hyperparameters for the models used in these experiments remain at their default settings.<sup>3</sup>

**Results** The experimental results for **Font Style** (**Plain**) are shown in Figure 3. These findings suggest that even minor adjustments to a specific font dimension, while keeping the textual content unchanged, can significantly impact the recognition performance of MLLMs. The results of the **Font** 

<sup>&</sup>lt;sup>1</sup>For detailed information, please refer to https://www.kaggle.com/datasets/kazanova/sentiment140

<sup>&</sup>lt;sup>2</sup>Figure 8 illustrates the five fonts separately.

<sup>&</sup>lt;sup>3</sup>For detailed configuration parameters, please refer to the website https://huggingface.co.



Figure 3: Results on font variations for text recognition using MLLMs with plain backgrounds.

Model	0-	9	0-9	Ð	0-:	9	0-	9	0-9	)
WIUUCI	ACC	NED	ACC	NED	ACC	NED	ACC	NED	ACC	NED
			E	English D	ataset					
Qwen2-VL-7B	25.90%	0.48	62.10%	0.56	45.10%	0.45	59.00%	0.31	57.90%	0.33
InternVL-1B	32.00%	0.34	28.70%	0.42	49.80%	0.28	48.50%	0.31	50.10%	0.28
InternVL-2B	23.70%	0.22	25.20%	0.28	45.10%	0.12	41.10%	0.13	42.10%	0.13
InternVL-4B	25.00%	0.31	13.20%	0.41	36.50%	0.28	32.10%	0.31	38.20%	0.27
MiniCPM-V2.6	43.20%	0.10	30.90%	0.23	73.60%	0.05	68.20%	0.05	73.20%	0.05
LLaVA-V1.6-Mistral-7B	0.00%	0.79	0.00%	0.83	0.00%	0.72	0.00%	0.77	0.00%	0.78
			C	hinese D	ataset					
Qwen2-VL-7B	75.30%	0.12	62.10%	0.13	80.00%	0.10	83.80%	0.08	84.80%	0.06
InternVL-1B	17.80%	0.42	18.00%	0.52	37.10%	0.36	34.20%	0.38	34.50%	0.38
InternVL-2B	21.70%	0.23	17.60%	0.29	32.10%	0.17	31.70%	0.19	42.10%	0.18
InternVL-4B	42.50%	0.11	44.60%	0.16	69.10%	0.07	63.00%	0.09	62.50%	0.08
MiniCPM-V2.6	41.00%	0.16	25.30%	0.21	65.40%	0.08	66.40%	0.08	64.40%	0.09

Table 2: The effect of font variations on scene text recognition using MLLMs for both English and Chinese datasets. Bold purple indicates the highest score across five fonts. The fonts from left to right are: *Debiao Pen Calligraphy*, *FZCYFW*, *Luxi Mono*, *Bold Oblique*, *Liugongquan Calligraphy*, and *IBMPlexSerif-LightItalic*.

Model	Norr	nal	Overl	ay	Highli	ght	Striketh	rough
	ACC	NED	ACC	NED	ACC	NED	ACC	NED
Qwen2-VL-7B	85.89%	0.12	50.62%-	0.23	86.38%+	0.11	60.18%-	0.61
MiniCPM-V2.6	90.09%	0.07	$50.62\%^{-1}$	0.18	$92.68\%^{+}$	0.06	$75.00\%^{-1}$	0.55
InternVL-1B	76.70%	0.18	39.55%-	0.32	79.11%+	0.16	52.90%-	0.62
InternVL-2B	77.86%	0.15	39.11%-	0.29	77.23%-	0.13	51.70%-	0.62
InternVL-4B	81.56%	0.12	43.97%-	0.25	$82.68\%^{+}$	0.10	64.38%-	0.57
LLaVA-V1.6-Mistral-7B	66.25%	0.65	33.26%-	0.73	$66.52\%^{+}$	0.63	35.89%-	0.84

Table 3: Evaluation of MLLMs for font semantics understanding.  $^-$  indicates a score decrease relative to Normal, while  $^+$  indicates an increase.

**Style (Scene)** experiment, shown in Table 2, indicate that font style variations can lead to inconsistent recognition outcomes in MLLMs, even with

397

398

399

constant textual content. Different MLLMs also have varying preferences for specific font styles.

402

The results for **Font Semantics** in Table 3 show



Figure 4: The impact of font style variations on sentiment analysis performance in multimodal large language models: A comparative study of cheerful (CF-A/B), horror (HF-A/B), and neutral (NF) fonts. Example Images in Appendix D.

that highlighting improves MLLMs' performance, while Overlay and Strikethrough reduce recognition scores compared to the Normal condition. This indicates that certain font-based modifications can obscure textual cues, leading to increased hallucination rates. The **Font Sentiment** results in Figure



Figure 5: Comparison of the ACC scores of the 6 MLLMs across 3 subtasks. Detailed ACC scores of all models on three subtasks can be found in Appendix E.

4 show that cheerful fonts (CF-A/B) boost positive recall, while horror fonts (HF-A/B) increase negative recall. The Neutral font (NF) yields the highest neutral recall for most models, except Qwen2-VL-7B and LLaVA-V1.6-Mistral-7B. These findings highlight that font variations influence sentiment classification in MLLMs, reinforcing the need for robustness against font-induced sentiment shifts.

#### 5.2 Benchmarking of MLLMs using FontHalu

We evaluated 30 models on the FontHalu benchmark, primarily selecting those with high scores on the OCRBenchmark as we considered them to be widely adopted and demonstrating superior KIE capabilities; the evaluated models encompassed both open-source and proprietary advanced MLLMs. A detailed list of model names and sources can be found in Appendix F. 422

423

424

425

426

427

428

429

430

431

432

433

434

435

436

437

438

439

440

441

442

443

444

445

446

447

448

449

450

451

452

453

454

455

456

**Settings** Following OCRbench (Liu et al., 2024e), the FontHalu benchmark adopts a questionanswering (QA) format, enabling efficient evaluation of models' capabilities. The querying prompts for each subtask are as follows: **VarRec**: "What's the content in the image? Only return the content." **SemAna**: "What's the {key} in the image? Only return the {key}." **SentAna**: "What emotions do the text in the picture express? Choose one from ['positive', 'negative', 'neutral']." All model parameters remain consistent with their official configurations, and API calls follow the same settings as the online versions to ensure reproducibility.

**Results** The evaluation results on the FontHalu benchmark are summarized in Table 4, where Qwen2-VL-7B achieves the highest score, followed by Qwen2.5-VL-7B in second place. Figure 5 visualizes the performance differences among the six models across the three sub-tasks. Notably, substantial variations are observed in the VarRec and SemAna sub-tasks. The Qwen series models demonstrate the best performance in VarRec, while GPT-40 achieves the highest accuracy in SemAna.

#### 5.3 Mitigation Strategies

We hypothesize that FontHalu in MLLMs arises due to insufficient exposure to diverse fonts during training, leading to poor generalization to unseen fonts. To test this, we use fine-tuning to mitigate FontHalu in this experiment.

**Settings** We first create a training dataset (*Same*) using 687 fonts that are included in the benchmark,

408

409

410

411

412

413

414

415

416

417

418

419

420

Name	ACC	Name	ACC
Qwen2-VL-7B	0.6649	Pixtral-12B	0.4872
Qwen2.5-VL-7B	0.6352	InternVL2-8B	0.4857
Gemini-1.5-Pro	0.6148	InternVL2-1B	0.4849
Gemini-Pro-Vision	0.5816	InternVL2-2B	0.4798
GPT-4o-2024-08-06	0.5661	CogVLM2-LLaMA3-19B-Chat	0.4616
MiniCPM-V-2.6	0.5601	LLaMA3.2-90B-Vision-Instruct	0.4359
GML-4V-9B	0.5599	Phi3.5-Vision-Instruct	0.4218
Phi-3-Vision-128k-Instruct	0.5527	Ovis1.5-LLaMA3-8B	0.3879
InternVL2.5-1B	0.5517	LLaVA-V1.6-Vicuna-7B-hf	0.3805
InternVL2.5-2B	0.5509	LLaMA3-LLaVA-Next-8B-hf	0.3647
Qwen2-VL-72B-Instruct	0.5499	LLaVA-V1.6-Mistral-7B	0.3597
InternVL2-4B	0.5306	LLaMA3.2-11B-Vision-Instruct	0.2866
InternVL2.5-8B	0.5259	DeepSeek-VL-7B-Chat	0.2339
InternVL2.5-4B	0.5251	InstructBLIP-Vicuna-7B	0.0349
MiniMax-01	0.4910	BLIP-VQA-Base	0.0284

Table 4: Overall performance of MLLMs on benchmark. The scores are arranged in descending order, from top to bottom, and from left to right.

ensuring the model encounters the same fonts dur-457 ing training and testing. Then we investigate the 458 model's ability to generalize to unseen fonts. We 459 create two additional datasets: Cluster-200 with 460 687 fonts from 200 clusters (excluding benchmark 461 fonts) and Cluster-10 (excluding benchmark fonts) 462 with 687 fonts from the top 10 clusters. Each font 463 464 in the training sets is used to generate 10 images, resulting in 6,870 samples per set. During evaluation, 465 we first test the models on the benchmark dataset. 466 Then, we assess performance on a more diverse test 467 set of 10,000 distinct fonts, excluding those used in 468 training. All test images are generated with Lorem 469 Ipsum text on a blank background<sup>4</sup>. 470

The results in Table 5 show that training Results 471 472 MLLMs with benchmark fonts improves performance, reducing hallucinations. Training with clus-473 tered fonts also helps, indicating the model's ability 474 to generalize across different fonts. However, the 475 Qwen2-VL-7B model's performance drops with 476 the Same or Cluster-200 datasets, which may be 477 attributed to the model's parameterization being un-478 suitable for uniform training samples, resulting in 479 overfitting to specific font styles. Although font cat-480 egory coverage during training has minimal impact 481 482 on benchmark performance, the table also shows that training with 200 font clusters generally results 483 in better performance on the test set compared to 484 485 training with only 10 clusters.

### 6 Conclusion

486

487

488

We conduct a comprehensive study on font-induced hallucinations in MLLMs, analyzing their manifes-

Model	<b>B-ACC</b>	T-ACC
LLaVA-V1.6-Mistral-7B	0.3588	0.0166
Same	0.3866↑	0.2344
Cluster-10	0.3876↑	<u>0.2357</u>
Cluster-200	0.3821↑	<b>0.2466</b>
MiniCPM-V-2.6	0.5576	0.2690
Same	0.5932↑	<u>0.3995</u>
Cluster-10	0.5923↑	0.3851
Cluster-200	0.6066↑	<b>0.4020</b>
Qwen2-VL-7B Same Cluster-10 Cluster-200	$0.6649 \\ 0.6595 \downarrow \\ 0.6671 \uparrow \\ 0.6561 \downarrow$	0.5905 <u>0.7244</u> 0.7097 <b>0.7245</b>

Table 5: Performance comparison of MLLMs on benchmark(B-ACC) and test datasets(T-ACC). Bold indicates the best method; underlined, the second-best.

tations across stylistic, semantic, and sentiment dimensions. Experimental results demonstrate that even subtle font variations can significantly impact model predictions. To systematically evaluate this phenomenon, we introduce FontHalu benchmark, the first benchmark specifically designed to assess MLLMs robustness against fonts variations. While certain models achieve strong performance on standard hallucination or OCR benchmarks, they exhibit relatively inferior results on FontHalu benchmark, revealing a previously unrecognized vulnerability in MLLMs' font processing capabilities. Furthermore, we propose a mitigation approach using font clustering to select representative typography for fine-tuning, reducing font-induced hallucinations. By systematically investigating FontHalu, we provide a foundation for developing more reliable, interpretable, and font-aware MLLMs.

489

<sup>&</sup>lt;sup>4</sup>Training parameters are detailed in Appendix G.

517

518

519

521

522

524

525

527

529

530

531

532

534

536

538

539

540

541

543

545

546

547

548

550

551

554

555

### Limitations

508 This study categorizes FontHalu into three distinct types, providing a useful framework for understand-509 ing font-related hallucinations. However, we ac-510 knowledge that this classification may not be exhaustive and could benefit from further exploration 512 513 of additional categories. As font-induced hallucinations are an emerging research area, further 514 studies are needed to refine and expand upon this 515 taxonomy. 516

Furthermore, although we cluster over 19,000 fonts and thoughtfully select 1,000 fonts for our benchmark, aiming to cover a wide range of typographic styles. However, we recognize that even with this selection, the full diversity of font styles may not be fully captured, and there is potential for expanding its scope. Font variations are virtually limitless, spanning different scripts, handwritten styles, and dynamically generated typefaces, which pose additional challenges not addressed in this study.

While we investigate fine-tuning as a mitigation strategy, our study does not explore alternative approaches such as in-context learning, prompt engineering, or reinforcement learning, which may offer more effective solutions. Additionally, we have not evaluated state-of-the-art hallucination mitigation methods on FontHalu. These limitations point to valuable directions for future research. Our primary contribution remains the systematic analysis of font-induced hallucinations in MLLMs and the establishment of the FontHalu benchmark.

#### References

- Chandranath Adak and Bidyut B. Chaudhuri. 2014. An approach of strike-through text identification from handwritten documents. In 2014 14th International Conference on Frontiers in Handwriting Recognition, pages 643–648.
- Aries Arditi and Jianna Cho. 2005. Serifs and font legibility. *Vision research*, 45(23):2926–2933.
- Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *CoRR*, abs/2404.18930.
- Mareike Bayer, Werner Sommer, and Annekathrin Schacht. 2012. Font size matters—emotion and attention in cortical responses to written words. *PloS one*, 7(5):e36042.
- Nik Bear Brown. 2024. The cognitive type project

- mapping typography to cognition. *ArXiv*, abs/2403.04087.

- Zhiyang Chen, Yousong Zhu, Yufei Zhan, Zhaowen Li, Chaoyang Zhao, Jinqiao Wang, and Ming Tang. 2023. Mitigating hallucination in visual language models with visual supervision. *CoRR*, abs/2311.16479.
- Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, Tianren Gao, Erlong Li, Kun Tang, Zhipeng Cao, Tong Zhou, Ao Liu, Xinrui Yan, Shuqi Mei, Jianguo Cao, Ziran Wang, and Chao Zheng. 2024. A survey on multimodal large language models for autonomous driving. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACVW 2024 - Workshops, Waikoloa, HI, USA, January 1-6, 2024*, pages 958–979. IEEE.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net.
- 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), pages 14375–14385.
- Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada, pages 18135–18143. AAAI Press.*
- Hongyu Hu, Jiyuan Zhang, Minyi Zhao, and Zhenbang Sun. 2023. CIEM: contrastive instruction evaluation method for better instruction tuning. *CoRR*, abs/2309.02301.
- Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. 2021. ICDAR2019 competition on scanned receipt OCR and information extraction. *CoRR*, abs/2103.10213.
- Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. FUNSD: A dataset for form understanding in noisy scanned documents. *CoRR*, abs/1905.13538.
- Chaoya Jiang, Hongrui Jia, Mengfan Dong, Wei Ye, Haiyang Xu, Ming Yan, Ji Zhang, and Shikun Zhang.

562

563

556

573

574

575

576

577

578

579

580

581

582

583

584

585

586

587

588

589

590

591

594

595

598

599

600

601

602

603

604

605

606

607

608

609

610

611

2024. Hal-eval: A universal and fine-grained hallu-

cination evaluation framework for large vision lan-

guage models. In Proceedings of the 32nd ACM

International Conference on Multimedia, pages 525-

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H.

Hoi. 2023a. BLIP-2: bootstrapping language-image

pre-training with frozen image encoders and large

language models. In International Conference on

Machine Learning, ICML 2023, 23-29 July 2023,

Honolulu, Hawaii, USA, volume 202 of Proceedings

of Machine Learning Research, pages 19730–19742.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang,

Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Eval-

uating object hallucination in large vision-language

models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Process-

ing, EMNLP 2023, Singapore, December 6-10, 2023,

pages 292-305. Association for Computational Lin-

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao,

and Ji-Rong Wen. 2023c. Evaluating object halluci-

nation in large vision-language models. In Proceed-

ings of the 2023 Conference on Empirical Methods in

Natural Language Processing, pages 292–305, Sin-

gapore. Association for Computational Linguistics.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser

Yacoob, and Lijuan Wang. 2024a. Mitigating hal-

lucination in large multi-modal models via robust instruction tuning. In The Twelfth International Con-

ference on Learning Representations, ICLR 2024,

Vienna, Austria, May 7-11, 2024. OpenReview.net.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae

Jiazhen Liu, Yuhan Fu, Ruobing Xie, Runquan Xie,

Xingwu Sun, Fengzong Lian, Zhanhui Kang, and

Xirong Li. 2024c. Phd: A chatgpt-prompted vi-

Yexin Liu, Zhengyang Liang, Yueze Wang, Muyang He,

Jian Li, and Bo Zhao. 2024d. Seeing clearly, answer-

ing incorrectly: A multimodal robustness benchmark

for evaluating mllms on leading questions. Preprint,

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang,

Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-

Lin Liu, Lianwen Jin, and Xiang Bai. 2024e. Ocr-

bench: on the hidden mystery of ocr in large multi-

modal models. Science China Information Sciences,

Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui

Chen, Chunyuan Li, Lianwen Jin, and Xiang Bai.

neural information processing systems, 36.

sual hallucination evaluation dataset.

arXiv:2403.11116.

arXiv:2406.10638.

Lee. 2024b. Visual instruction tuning. Advances in

617

534.

PMLR.

guistics.

- 618
- 619

- 624 625

- 634
- 636
- 640

- 648

664

- 665

- - 2023. On the hidden mystery of OCR in large multimodal models. CoRR, abs/2305.07895.

67(12).

Stuart P. Lloyd. 1982. Least squares quantization in PCM. IEEE Trans. Inf. Theory, 28(2):129-136.

670

671

672

673

674

675

676

677

678

679

680

681

682

683

684

685

686

687

688

689

690

691

692

693

694

695

696

697

698

699

700

701

702

703

704

705

706

707

708

709

710

711

713

714

715

716

717

718

719

720

721

722

723

724

- Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. 2023. Negative object presence evaluation (NOPE) to measure object hallucination in vision-language models. CoRR, abs/2310.05338.
- Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. Docvqa: A dataset for VQA on document images. In IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021, pages 2199–2208. IEEE.
- Leland McInnes and John Healy. 2018. UMAP: uniform manifold approximation and projection for dimension reduction. CoRR, abs/1802.03426.
- Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the* 2018 Conference on Empirical Methods in Natural Language Processing, pages 4035-4045, Brussels, Belgium. Association for Computational Linguistics.
- Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA models that can read. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, pages 8317-8326. Computer Vision Foundation / IEEE.
- 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. Qwen2vl: Enhancing vision-language model's perception of the world at any resolution. CoRR, abs/2409.12191.
- Jen-Her Wu and Yufei Yuan. 2003. Improving searching and reading performance: the effect of highlighting and text color coding. Information & Management, 40(7):617-637.
- Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800.
- Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. 2024. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024, pages 12944-12953. IEEE.
- Shuo Zhang, Pengjiang Wang, and Wenjun Hou. 2018. Research on font emotion based on semantic difference method. In International Conference on Human Centered Computing, pages 304-313. Springer.
- 10

Preprint,

725

- 732 733 734 735
- 738
- 740

741

742

743

744

746

747

748

750

751

752

754

758

759

767

770

772

773

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. Analyzing and mitigating object hallucination in large vision-language models. In The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net.

### A Details on the Font Clustering Algorithm

To effectively organize and analyze the highdimensional feature vectors of fonts, we adopt a two-step approach comprising dimensionality reduction using Uniform Manifold Approximation and Projection (UMAP) and clustering using Kmeans. This methodology facilitates the identification of representative fonts by leveraging the complementary strengths of UMAP's non-linear dimensionality reduction and K-means' centroidbased clustering.

High-dimensional feature vectors are often challenging to visualize and cluster due to the "curse of dimensionality" and the complex relationships between features. To address this, we employ Uniform Manifold Approximation and Projection (UMAP)(McInnes and Healy, 2018), a state-of-theart non-linear dimensionality reduction technique known for preserving both local and global structures of high-dimensional data (McInnes and Healy, 2018). Given a set of high-dimensional vectors representing font features:

$$X = \{x_1, x_2, \dots, x_n\}, \quad x_i \in \mathbb{R}^d,$$

where n denotes the number of fonts and d represents the dimensionality of the feature space, UMAP projects these vectors into a lowerdimensional space  $\mathbb{R}^m$  while preserving the topological structure:

$$f: \mathbb{R}^d \to \mathbb{R}^m, \quad Z = f(X),$$

where  $Z = \{z_1, z_2, ..., z_n\}$  and  $m \ll d$ . UMAP constructs a weighted k-nearest neighbor graph to model the local relationships between highdimensional points, optimizing the following crossentropy objective:

$$C = -\sum_{i \neq j} \left[ p_{ij} \log q_{ij} + (1 - p_{ij}) \log(1 - q_{ij}) \right],$$

where  $p_{ij}$  is the conditional probability of  $x_i$  being close to  $x_j$  in the high-dimensional space, estimated using a Gaussian kernel.  $q_{ij}$  is the corresponding probability in the lower-dimensional space, parameterized as a Student's t-distribution to better capture local structures. This optimization preserves the local neighborhood continuity while maintaining the global data manifold, thus ensuring that similar fonts are embedded closely together in the reduced space.

Following dimensionality reduction, the transformed vectors are clustered using K-means, a widely-used centroid-based clustering algorithm that minimizes the within-cluster variance. Specifically, the vectors are partitioned into K = 200clusters to capture the diverse stylistic variations present in the font dataset. The objective function of K-means is defined as follows:

$$I = \sum_{k=1}^{K} \sum_{i \in C_k} \|z_i - \mu_k\|^2,$$
792

774

775

776

777

778

779

780

781

782

783

784

785

786

787

788

789

790

791

793

794

795

796

797

799

800

801

803

804

806

807

808

809

810

811

812

813

814

815

816

817

where: $C_k$  denotes the set of vectors assigned to cluster k.  $\mu_k$  is the centroid of cluster  $C_k$ , calculated as:

$$\mu_k = \frac{1}{|C_k|} \sum_{i \in C_k} z_i.$$

 $\|\cdot\|$  represents the Euclidean distance. The K-means algorithm alternates between the following two steps until convergence:

1. Assignment Step: Assign each vector to the nearest centroid:

$$C_k = \left\{ z_i : \|z_i - \mu_k\|^2 \le \|z_i - \mu_j\|^2, \forall j, 1 \le j \le K \right\}$$

2. Update Step: Recalculate the centroids as the mean of all vectors in each cluster:

$$\mu_k = \frac{1}{|C_k|} \sum_{i \in C_k} z_i \tag{805}$$

This iterative procedure continues until convergence, yielding 200 distinct clusters. From each cluster, five representative fonts are selected based on their proximity to the cluster centroid, ensuring that the chosen fonts are central and representative of their respective clusters:

$$F_k = \arg\min_{x \in C_k} \|z_i - \mu_k\|, \quad |F_k| = 5,$$

where  $F_k$  denotes the set of five representative fonts from cluster k. This process yields a total of **1,000** candidate fonts, providing a diverse yet concise benchmark set. Finally, the result of kmeans is shown in the figure6.



Figure 6: Font clustering result diagram.

### 818 B Annotation and Filtering Criteria

#### 9 The **annotation criteria** are as follows:

820

821

824

825

828

829

831

832

- 1. The bounding box (bbox) should ensure that the corresponding keyword is clearly visible, facilitating subsequent queries about the content within the bbox based on the keyword.
- 2. The bounding box should align as closely as possible with the original content's location, avoiding discrepancies with the actual key content.
- 3. When annotating, the size of the bounding box should match the actual content's dimensions, avoiding excessive cropping or expansion.
- 4. The bounding box should not obscure any crucial information, ensuring that all key content is fully presented.

#### 4 The **filtering criteria** are as follows:

1. Images with truncated content are filtered out.

- Images with garbled content, where the corresponding font fails to render ASCII characters, are excluded. Images where the content is unclear or fuzzy (i.e., human-readable keywords are difficult to discern) are removed.
- 3. During the verification process, three reviewers assess the image, and it is only discarded if all three reviewers agree that the content is unreadable.

841

842

843

844

845

846

847

848

# C Evaluation Metrics

Here,  $y_i$  denotes the expected answer with spaces removed and  $\hat{y}_i$  denotes the generated answer with spaces removed:

1. ACC: For each sample: 849

$$Score_i = 1$$
 if  $y_i$  in  $\hat{y}_i$ , else 0. 850

$$Acc(Y, \hat{Y}) = \frac{1}{N} \sum_{i=1}^{N} Score_i(y_i, \hat{y}_i).$$
852

2. NED:

- 855

- 864

- 873 874 876
- 878
- 881

893

897 898

901

 $\text{NED} = \frac{1}{N} \sum_{i=1}^{N} \frac{\text{EditDistance}(y_i, \hat{y}_i)}{\max(|y_i|, |\hat{y}_i|)}$ 

#### D **Detailed Experiment on Font-based** Hallucination

This section primarily provides supplementary details for the experiments in Section 5.1.

In the Font Style (Plain) experiment, Roboto, a widely used font from Google Fonts, serves as the base font. To isolate the effect of each dimension. only one dimension is varied at a time, while the others remain at their default values(e.g., size of 20, weight of 400, alignment as left, spacing as 0em, slant as 0, and line-height as 1). The canvas size is set to 1024x256 pixels. We begin with a set of 200 images, each featuring a blank background to eliminate layout interference. Each image contains three lines of text: the first line consists of random numbers, the second line contains random letters, and the third line features a sentence. For each dimension under investigation, variations are applied to these 200 base images. The multimodal large model's task is to recognize and extract all text content from these images. An example can be seen in the top-left corner of the "font variations" box in Figure 7.

In the Font Style (Scene) experiment, we select five representative fonts, including both English and Chinese styles. The experimental scenarios cover common numerical applications, such as tracking numbers, ID cards, invoices, and transportation permits. We first collect data from the aforementioned scenarios, then annotate the data with bounding boxes (bbox), removing the original content within the boxes and replacing it with new content, restricting it to numerical data to eliminate language-related factors. The task is to have MLLMs recognize the content within these boxes. Finally, we constructed a dataset consisting of 2,000 Chinese samples and 2,000 English samples.

In the Font Semantics experiment, similar to the Font Style (Scene) experiment, we draw specific content on images from various scenarios and task MLLMs with recognizing the content. The key difference is that the applied content undergoes four random format transformations: normal, strikethrough, overlay, and highlight. We select the Roboto font and generate 2,000 data samples for each format to conduct our exploration.

In **Font Sentiment** experiment, We select 1,000 samples from the test set of the Sentiment140 dataset and render the text using three font styles: a visually neutral font, a horror-themed font, and a visually cheerful font. Those fonts can be seen in Figure 8. The images are then evaluated by MLLMs to determine which emotion is conveyed by texts in the images.

902

903

904

905

906

907

908

909

910

911

912

913

914

915

916

917

918

919

920

921

922

923

924

925

926

927

928

929

930

931

932

933

934

935

936

937

938

939

940

941

942

943

944

945

946

947

948

#### **Detailed Scores on Subtasks of** Ε **Benchmark**

The table 6 shows the scores of each subset in the benchmark, including VarRec, SemAna, and SentAna. The scores of each subset can be seen from the table. It can be seen that the highest score of VarRec is the model Owen2-VL-7B, with an accuracy of 62.80%. There are still many models that do not have the ability to recognize long outof-order texts, with an accuracy rate below 10%.

In the SemAna subset, the main tasks are KIE recognition tasks in the context of documents and invoices. The length of the content to be recognized is not as long as that of the VarRec dataset, so the overall recognition difficulty is relatively low, and the weaker models also have a certain accuracy rate. The highest accuracy rate is also the Qwen2-VL-7B model, with an accuracy rate of 68.19%, beating many models with larger parameters, which also indirectly shows that models with larger parameters are not necessarily more robust to fonts in KIE recognition tasks.

#### **Benchmark Models** F

Table 8 lists all models evaluated in this paper, including the ways in which the institution to which the model belongs has acquired it.

#### G **Training Configuration Details**

We conducted parameter-efficient fine-tuning using Low-Rank Adaptation (LoRA) with a rank value of 8 on three vision-language architectures: LLaVA-1.6-Mistral-7B, MiniCPM-V-2.6, and Qwen2-VL-7B. To achieve comprehensive adaptation of multimodal representations, LoRA injections were applied to all linear layers-including attention mechanisms and feedforward networks-with a base learning rate of  $1 \times 10^{-5}$  regulated by cosine decay scheduling and a 10% warmup phase. The training process employed gradient accumulation over 8 steps per



Figure 7: Example images from the font-based hallucination experiment in Section 5.1. [Q] represents the question, and [G] represents the ground truth.



Figure 8: The fonts used in the images from top to bottom are Neutral Font, Cheerful Font A, Cheerful Font B, Horror-themed Font A, and Horror-themed Font B.

device coupled with **FP16** mixed-precision arithmetic. The experiments are all conducted on the A100 for both inference and fine-tuning.

#### H Evaluation Samples

 Here are some benchmark recognition examples 9, 10, 11, 12. Each example shows the answers of 4 models, namely Qwen2-VL-72B, Qwen2-VL-7B, Minicpm-V-2.6 and GPT40 models.

Model	VarRec	SemAna	SentAna
Qwen2-VL-7B	0.6280	0.6819	0.5690
Qwen2.5-VL-7B	0.4560	0.6622	0.6120
Gemini-Pro-1.5	0.2680	0.6718	0.5315
Gemini-Pro-Vision	0.0480	0.6548	0.5696
GPT-40-2024-08-06	0.4280	0.5717	0.6702
MiniCPM-V-2.6	0.3640	0.5876	0.5510
Glm-4v-9B	0.3440	0.5850	0.5918
Phi-3-Vision-128k-Instruct	0.1080	0.6244	0.4577
InternVL2.5-1B	0.2160	0.6095	0.4504
InternVL2.5-2B	0.2120	0.6137	0.4136
Qwen2-VL-72B-Instruct	0.3760	0.5637	0.6259
InternVL2-4B	0.1880	0.5855	0.4597
InternVL2.5-8B	0.2080	0.5690	0.5227
InternVL2.5-4B	0.2240	0.5738	0.4599
Minimax-01	0.2533	0.5431	0.3305
Pixtral-12B	0.2960	0.4997	0.5916
InternVL2-8B	0.1480	0.5328	0.4726
InternVL2-1B	0.1320	0.5605	0.2608
InternVL2-2B	0.1160	0.5429	0.3669
CogVLM2-LLaMA3-19B-Chat	0.0840	0.5003	0.5574
LLaMA3.2-90B-Vision-Instruct	0.0840	0.4758	0.4947
Phi3.5-Vision-Instruct	0.1200	0.4513	0.5091
Ovis1.5-LLaMA3-8B	0.1200	0.4156	0.4541
LLaVA-V1.6-Vicuna-7B-hf	0.0120	0.4182	0.4744
LLaMA3-LLaVA-Next-8B-hf	0.0240	0.3953	0.4856
LLaVA-V1.6-Mistral-7B	0.0160	0.3937	0.4567
LLaMA3.2-11B-Vision-Instruct	0.1440	0.3031	0.3077
Deepseek-VL-7B-Chat	0.0760	0.2339	0.4026
InstructBLIP-Vicuna-7B	0.0000	0.0027	0.3285
BLIP-VQA-Base	0.0000	0.0000	0.2844

Table 6: Detailed ACC Scores of MLLMs on Three Subtasks

Model	VarRec	SemAna
Qwen2-VL-7B	0.0663	0.2048
Qwen2.5-VL-7B	0.1396	0.1950
Gemini-Pro-1.5	0.1040	0.2061
Gemini-Pro-Vision	0.3887	0.2468
GPT-4o-2024-08-06	0.0920	0.2821
MiniCPM-V-2.6	0.1017	0.2706
Glm-4v-9B	0.1942	0.7000
Phi-3-Vision-128k-Instruct	0.8270	0.2897
InternVL2.5-1B	0.1776	0.2498
InternVL2.5-2B	0.1777	0.2468
Qwen2-VL-72B-Instruct	0.2001	0.2880
InternVL2-4B	0.2330	0.2852
InternVL2.5-8B	0.1619	0.2611
InternVL2.5-4B	0.1731	0.2603
Minimax-01	0.1225	0.3095
Pixtral-12B	0.1302	0.3307
InternVL2-8B	0.1910	0.4544
InternVL2-1B	0.2390	0.3386
InternVL2-2B	0.2965	0.2826
CogVLM2-LLaMA3-19B-Chat	0.7858	0.8073
LLaMA3.2-90B-Vision-Instruct	0.8038	0.5234
Phi3.5-Vision-Instruct	0.8006	0.4282
Ovis1.5-LLaMA3-8B	0.1671	0.3560
LLaVA-V1.6-Vicuna-7B-hf	0.5810	0.3567
LLaMA3-LLaVA-Next-8B-hf	0.7222	0.4786
LLaVA-V1.6-Mistral-7B	0.8612	0.7885
LLaMA3.2-11B-Vision-Instruct	0.6980	0.6670
Deepseek-VL-7B-Chat	0.5890	0.9248
InstructBLIP-Vicuna-7B	0.9314	0.9122
BLIP-VQA-Base	0.9565	0.9314
~		

Table 7: Detailed edit distance Scores of MLLMs on Two Subtasks

Model	Creator	Access
Qwen2-VL-7B	Alibaba	Weights
Qwen2.5-VL-7B	Alibaba	Weights
Gemini-Pro-1.5	Google	API
Gemini-Pro-Vision	Google	API
GPT-40	OpenAI	API
MiniCPM-V-2.6	OpenBMB	Weights
Glm-4v-9B	Zhipu AI	Weights
Phi-3-Vision-128k-Instruct	Microsoft	Weights
InternVL2.5-1B	Shanghai AI Lab	Weights
InternVL2.5-2B	Shanghai AI Lab	Weights
Qwen2-VL-72B-Instruct	Alibaba	API
InternVL2-4B	Shanghai AI Lab	Weights
InternVL2.5-8B	Shanghai AI Lab	Weights
InternVL2.5-4B	Shanghai AI Lab	Weights
Minimax-01	Minimax	API
Pixtral-12B	Mistral AI	API
InternVL2-8B	Shanghai AI Lab	Weights
InternVL2-1B	Shanghai AI Lab	Weights
InternVL2-2B	Shanghai AI Lab	Weights
CogVLM2-LLaMA3-19B-Chat	Zhipu AI	Weights
LLaMA3.2-90B-Vision-Instruct	Meta	API
Phi3.5-Vision-Instruct	Microsoft	Weights
Ovis1.5-LLaMA3-8B	Alibaba	Weights
LLaVA-V1.6-Vicuna-7B-hf	UW-Madison	Weights
LLaMA3-LLaVA-Next-8B-hf	UW-Madison	Weights
LLaVA-V1.6-Mistral-7B	UW-Madison	Weights
LLaMA3.2-11B-Vision-Instruct	Meta	API
Deepseek-VL-7B-Chat	DeepSeek	Weights
InstructBLIP-Vicuna-7B	Salesforce	Weights
BLIP-VQA-Base	Salesforce	Weights

Table 8: Models evaluated in this paper

22570774 QIAUIQU TEVIL OREMLDO EEUQI	۲
🔅 Qwen2VL-7B: 22570774 QIAUIQU TEVIL OREMILDO EEUQN	X
🐼 Qwen2VL-72B: 22570774 QIAUIQU TEVIL OREMIDO EEUQN	X
MiniCPM-V2.6: 22570774 QIAUIQU TEVIL OREMLDO EEUQN	$\checkmark$
GPT4o: 225707774\nçQIAUIQUTEVILOREMLDOU EEUQN	X
16451559 etam mdoi UMQNMUA	
	r
☞ Qwen2VL-7B: 16451559 etam mdoi UMQNMUA 🗸	,
MiniCPM-V2.6: 16451559 etam mdoi UMQNNUA X	,
GPT4o: 16451559 etam mdoi UMQNMUA √	
RAPETMO 91331392 IAIPESD	7
🐼 Qwen2VL-72B: Rapetmo 913331392 lalpcsdi	X
Image: Weight of the second state of the second stat	$\checkmark$
MiniCPM-V2.6: RAPETMO 31331392 IAIPESDI	X
GPT4o: KAPETMO 91331982 TAIROSVI	X

Figure 9: Some examples from the VarRec dataset, where the red font shows the inconsistency between the model's answer and the correct answer. As shown in the figure, due to changes in font style, there may be some recognition illusions, such as the presence of confusing characters, such as the 'L' character in example 1, which the model may recognize as 'I', and the 'M' character in example 2, which may be recognized as 'N'.

()	Key: Name right answer: JOSHUA WILLIAMSON
n and a	🚱 Qwen2VL-72B: Julie Davis Joshua Williamson 🛛 🗙
Contraction Company	☞ Qwen2VL-7B: JULIE DAVIS JOSHUA WILLIAMSON X
Name JULIE DAVIS JOSHUA WILLIAMSON	MiniCPM-V2.6: JULIE DAVIS JOSHUA WILLIAMSON X
Mailing Addless         Kf # '1 Dolf 29           V1V         Chy (1) by (1) by (2)           State (1)         Chy (2)           Zip Code         3D479	ஞ GPT40: JOSHUA WILLIAMSON √
Home Telephone 63261571544	
Cell Telephone E-Mail Address	Key: City right answer: New Aaronton
Owned Barns Type Model Fuel Number Mapufacturer (Box or frack) Visia (MG, LPG, OI) Of Barns	🐼 Qwen2VL-72B: New London X
Jong the 1304 80-46 276 23 5	Qwen2VL-7B: New Arenton X
Land (Acres) Production 1999 Orop Est. 2000 Orop (Pounds) (Pounds)	MiniCPM-V2.6: Newark X
Owned <u>670</u> Owned <u>219,000</u> Owned <u>229,800</u> Leased <u>600</u> Leased Lessed	GPT40: New Agawam X
Irrigation Yes No	Q
When will your farm be available for a visit from RJR representatives?	Key: TELEPHONE right answer: 63261571544
comments: Ready to terre from year.	w Qwen2VL-72B: 63261571544 √
LAW SUIT	💊 😪 Qwen2VL-7B: 63261571544 🗸 ✓
Source: https://www.industrydeaumente.weafc.dw/does/treyw8284	MiniCPM-V2.6: 63261571544 🗸
	GGPT40: 63261571544 ✓

Figure 10: An example from the SemAna dataset. In this example, not only will various fonts be introduced, but also some semantic tasks such as overlay, highlight, and stricthrough will be added. In this case, there are issues such as confusion between uppercase and lowercase, incorrect recognition of the key as a whole, and errors in character-level recognition of the key, etc.



Figure 11: Another example from the SemAna dataset: In the name field recognition task in this example, many models identify incorrect fields and do not account for the interference caused by strikethrough.

GTONFOOLERYSZ YES. I M THE REMUS LUPIN OF WEREBENRS. JUST WITCH OUT NT THE NEXT FULL MOON. I MAY ONCE NOWN NITEMPT TO STELL YOUR HONEY
What emotions do the text in the picture express? Choose one from ['positive', 'negative', 'neutral'] right answer: positive
Qwen2VL-72B: The text in the picture expresses a sense of mischief and playfulness, which can be categorized as 'positive' emotions.
Qwen2VL-7B: The text in the picture expresses a negative emotion. X
MiniCPM-V2.6: Positive. 🗸
GPT40: Neutral. X
"Bok squad announcenery will be four row for the Australasian four, open training session for public at orlando stadium at 15:00 "Thursday"
What emotions do the text in the picture express? Choose one from ['positive', 'negative', 'neutral'] right answer:neutral
Qwen2VL-72B: The text in the picture expresses a positive emotion. The announcement of a Bok squad and an open training session for the public at an iconic stadium like Orlando Stadium suggests excitement and anticipation for the event.
5 Qwen2VL-7B: The text in the picture expresses a positive emotion. $X$
MiniCPM-V2.6: The text in the picture expresses a neutral emotion. The content of the message is informational, announcing an event without conveying any particular positive or negative sentiment $\checkmark$
$\bigcirc$ GPT40: The text in the picture expresses a neutral emotion. $\checkmark$

Figure 12: An example from the SentAna dataset. The text content in the example above is "@ TomFoolery23 Yes, I am the Remus Lupin of werebiers. Just watch out at the next full moon. I may once again attend to steam your honey...". This sentence references the character Remus Lupin from the Harry Potter series, who is a werewolf. Here, someone humbly refers to themselves as the 'werebier version of Lupin' and enjoyable mentions that they might try to steam honey during the next full moon. Clearly, it's a joke with a question. However, some models has identified a neutral, or even negative sentiment. The text content in the example below is "Book special announcement will be tomorrow for the Australian tour, open training session for public at Orlando stadium at 15:00 Thursday." This is an activity description without any specific emotions, and the answer is neutral, but some models answer with positive emotions.