

# 000 001 002 003 004 005 006 007 008 009 010 A HIGH QUALITY DATASET AND RELIABLE EVALUA- TION FOR INTERLEAVED IMAGE-TEXT GENERATION

005  
006  
007  
008  
009  
010  
011  
012  
013  
014  
015  
016  
017  
018  
019  
020  
021  
022  
023  
024  
025  
026  
027  
028  
029  
030  
031  
032  
033  
034  
035  
036  
037  
038  
039  
040  
041  
042  
043  
044  
045  
046  
047  
048  
049  
050  
051  
052  
053  
Anonymous authors  
Paper under double-blind review

## ABSTRACT

Recent advancements in Large Multimodal Models (LMMs) have significantly improved multimodal understanding and generation. However, these models still struggle to generate tightly interleaved image-text outputs, primarily due to the limited scale, quality and instructional richness of current training datasets. To address this, we introduce **InterSyn**, a dataset that features: (1) large scale, comprising 1.8M multimodal samples; (2) high quality, supported by our proposed **Self-Evaluation with Iterative Refinement (SEIR)** method for rigorous automated quality refinement; (3) rich instructional diversity, ensured through diverse well-designed question templates, based on human preferences and covering a 3500-topic hierarchy. These characteristics make InterSyn particularly well-suited for training LMMs in interactive image–text generation capabilities. To evaluate the capabilities, we propose **SynJudge**, a reliable automatic evaluator that aligns closely with human judge and outputs four interpretable scores: Text Content Completeness (TCC), Image Content Completeness (ICC), Image Quality (IQ), and Image–Text Synergy (ITS). These scores are complementary, covering both content and quality as well as cross-modal interaction, thereby forming a comprehensive evaluation framework. Experimental results on InterSyn subsets of up to 200K samples show that 25K–50K already yield substantial improvements, while scaling to 100K/200K brings further gains in TCC, ICC, and especially ITS, highlighting InterSyn’s: (1) scalability, as performance consistently improves with more data; (2) efficiency, as significant gains are achievable even with smaller subsets, making it accessible to researchers with varying computational resources.

## 1 INTRODUCTION

Multimodal understanding and generation are critical capabilities toward artificial general intelligence. In the past two years, multimodal large language models (MLLMs) (Liu et al., 2023; Chen et al., 2024c; Wang et al., 2024a) have shown remarkable performance in multimodal understanding and even surpassed humans in some areas, while we have also seen many impressive advances in high quality image generation (Esser et al., 2024b; Betker et al., 2023). However, these models are often limited to generating either text or image outputs in isolation, while real-world scenarios typically require tightly interleaved multimodal outputs.

Recently, pioneer unified LMMs, such as Janus-Pro (Chen et al., 2025b), have shown great potential. However, they struggle to generate instruction-following interleaved image-text outputs, manifesting issues such as semantic drift, low image–text synergy, and poor image quality.

The main challenges lie in the limited scale, quality, and instructional richness of existing datasets. Even with existing datasets (Zhu et al., 2023; Laurençon et al., 2023; Chen et al., 2024a;b; Xu et al., 2024), these challenges remain due to their critical limitations: (1) **Limited scale**: Focus on narrow tasks and typically contain no more than tens of thousands of samples, limiting their applicability to broader real-world scenarios; (2) **Unstable quality**: Built on web-crawled sources (Yang et al., 2021; Laurençon et al., 2023) or reused corpora (Zhou et al., 2018; Zhu et al., 2023) with inconsistent quality and lack standardized quality control mechanisms; (3) **Low interaction complexity**: Rely on static documents or single-turn prompts, thus failing to capture the contextual richness and interleaved structure of authentic human conversations.

To address the above issues, we introduce **InterSyn**—the first fully automated, high quality, large scale dataset for instruction-following, multi-turn question answering with interleaved image–text

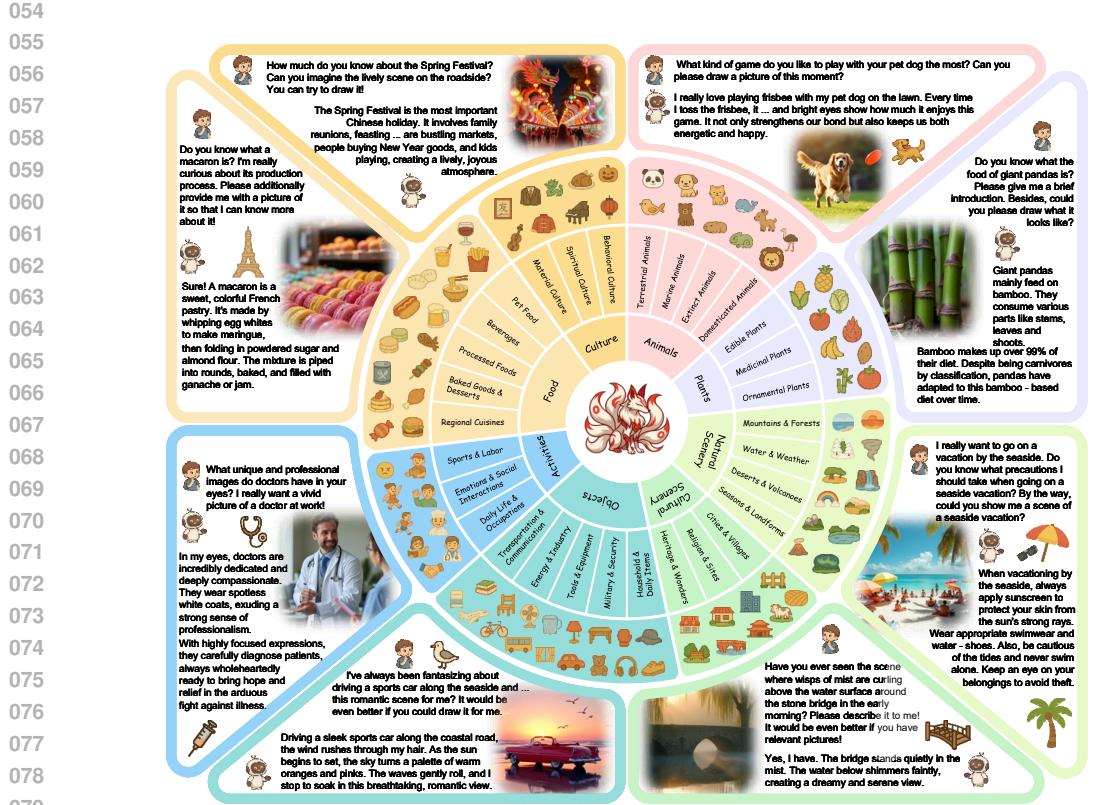


Figure 1: InterSyn: Topic hierarchy and interleaved question answering samples

responses. With 1.8M single-turn and 50,000 multi-turn dialogues across 8 domains and 3,500 topics (as shown in Figure 1), InterSyn provides extensive coverage for diverse real-world scenarios. InterSyn's quality is further enhanced by our **Self-Evaluation with Iterative Refinement (SEIR)** method, which embeds self-checking and feedback loops into each generation step, enhancing semantic completeness and cross-modal synergy. To ensure rich instructional diversity, we extract human-like query styles through question templates that capture varied linguistic structures, supported by a rich topic hierarchy for diverse, instruction-driven dialogues.

To evaluate interleaved image–text generation, several related benchmarks (An et al., 2024; Zhou et al., 2024; Chen et al., 2025a; Xia et al., 2024; Liu et al., 2024) have been proposed, but they still suffer from the following limitations: (1) **Limited domain and scale**: small, task-specific sets cannot cover realistic multi-turn dialogue needs; (2) **Costly manual evaluation**: accurate assessment still hinges on human, whose expense and delay hinder large scale, rapid benchmarking; (3) **Weak alignment with human preference**: current automatic metrics significantly diverge from human judgments on fine-grained multimodal reasoning; (4) **Narrow evaluation scope**: emphasise surface correctness while overlooking synergy and overall answer quality.

Therefore, we propose **SynJudge**, a reliable and comprehensive judge model for evaluating interleaved image–text generation with high alignment to human judgment. SynJudge provides interpretable, quantitative feedback across four key dimensions: text content completeness (TCC), image content completeness (ICC), image quality (IQ), and image–text synergy (ITS). Unlike traditional image–text consistency metrics, the ITS metric focuses on rewarding tight, complementary alignment between the textual and visual modalities while penalizing redundancy.

To validate our contributions, we conduct a series of experiments. First, the effectiveness of our SEIR method is demonstrated, showing substantial quality improvements over a non-refined baseline. Additionally, comparisons with several existing models show that SEIR consistently outperforms them across all evaluation metrics. Next, we validate SynJudge, which exhibits the strongest alignment with human judgment, showing a significantly smaller deviation compared to zero-shot MLLM evaluators. Finally, we validate the utility of InterSyn by fine-tuning models on randomly sampled subsets, up to 200K examples. Results reveal significant performance gains with as few as

108 25K/50K samples, underscoring the dataset’s high data density and efficiency. These results confirm  
 109 that our methods and dataset contribute to improved multimodal model performance.  
 110

111 Our main contributions are summarized as follows: (1) We present **InterSyn**, a large scale dataset of  
 112 1.8M high-fidelity samples, distinguished by its instructionally rich and complex dialogues that span  
 113 over 3,500 topics; (2) We propose **SEIR**, a method that ensures high quality data generation across  
 114 refinement steps with minimal manual effort; (3) We introduce **SynJudge**, a multi-dimensional  
 115 evaluation model for scoring interleaved outputs, enabling fine-grained assessment and effective  
 116 feedback for model improvement; (4) We conduct comprehensive experiments demonstrating that  
 117 InterSyn substantially enhances LMM performance in instruction alignment, image-text synergy,  
 118 and multi-turn reasoning, contributing to the advancement of unified multimodal systems.  
 119

## 2 RELATED WORK

120 **Models for Interleaved Image-Text Generation.** Recent advances in MLLMs, such as  
 121 Flamingo (Alayrac et al., 2022), InternVL (Chen et al., 2024d), and Qwen-VL (Wang et al., 2024a),  
 122 have substantially improved multimodal understanding. Meanwhile, diffusion models (Ramesh  
 123 et al., 2022; Betker et al., 2023; Esser et al., 2024a) achieve strong visual generation performance.  
 124 To unify understanding and generation, models such as MiniGPT-5 (Zheng et al., 2023) and Show-  
 125 o (Xie et al., 2024) combine autoregressive text generation with diffusion-based image synthesis.  
 126 More recent efforts (Team, 2024; Wu et al., 2024; Wang et al., 2024b; Chern et al., 2024) adopt uni-  
 127 fied autoregressive frameworks for interleaved generation. However, lacking targeted, high quality  
 128 training data, these models are not explicitly optimized for instruction-following and often struggle  
 129 to maintain coherence and cross-modal consistency—a gap InterSyn is specifically designed to fill.  
 130

131 **Datasets for Interleaved Image-Text Generation.** High quality interleaved image-text data is cru-  
 132 cial for training multimodal models. Existing large scale datasets like MMC4 (Zhu et al., 2023),  
 133 OBELICS (Laurençon et al., 2023), and CoMM (Chen et al., 2024b) are primarily document-level  
 134 corpora constructed from web sources, but often suffer from noise, weak alignment and low inter-  
 135 action intensity. Several benchmarks, such as OpenLEAF (An et al., 2024), InterleavedBench (Liu  
 136 et al., 2024), and OpenING (Zhou et al., 2024), focus on specific tasks. LeafInstruct (Xu et al.,  
 137 2024) constructs an interleaved image-text dataset by filtering samples from existing corpora (Zhu  
 138 et al., 2023; Huang et al., 2016; Zhou et al., 2018). However, both benchmarks and datasets remain  
 139 limited in scale and instructional diversity. To this end, we introduce InterSyn, a large scale, high  
 140 quality dataset with diverse, multi-turn dialogues and automated refinement.  
 141

142 **Evaluation for Interleaved Image-Text Outputs** Early multimodal evaluation metrics indepen-  
 143 dently assessed text quality (Papineni et al., 2002; Lin, 2004) and image quality (Heusel et al., 2017;  
 144 Salimans et al., 2016). Subsequent metrics (Hessel et al., 2021; Li et al., 2023; Lin et al., 2024;  
 145 Chen et al., 2023; Lu et al., 2024) targeted image-text consistency, yet still inadequately evaluated  
 146 the quality of interleaved outputs. More recent efforts, including InterleavedEval (Xu et al., 2024)  
 147 and CoMM (Chen et al., 2024b), leveraged MLLMs for holistic assessment, but often exhibit mis-  
 148 alignment with human judgment. OpenING (Zhou et al., 2024) proposed IntJudge for pairwise  
 149 comparisons, but it lacks fine-grained, quantitative scoring for individual responses, limiting its ap-  
 150 plicability for model training and refinement. In contrast, the proposed SynJudge provides a more  
 151 comprehensive evaluation by assessing both content completeness and modality synergy, aligning  
 152 more closely with human judgment.  
 153

## 3 INTERSYN DATASET AND SYNJUDGE EVALUATOR

### 3.1 OVERVIEW

154 In this section, we present a comprehensive framework for both the construction and evaluation of  
 155 interleaved image-text generation. First, we describe the InterSyn dataset construction pipeline, de-  
 156 tailing the preparatory work and our proposed **SEIR** method (illustrated in Figure 2). Subsequently,  
 157 we introduce **SynJudge**, a specialized evaluator designed to assess the quality of interleaved outputs,  
 158 along with the specific metrics and benchmarks established to validate model performance.  
 159

### 3.2 DATASET PREPARATORY WORK

160 InterSyn’s preparatory work involves five major stages:

161 **Question Collection.** We recruited 25 participants, each providing 40 questions drawn from natural  
 162 conversational scenarios, resulting in a total of 1,000 questions.

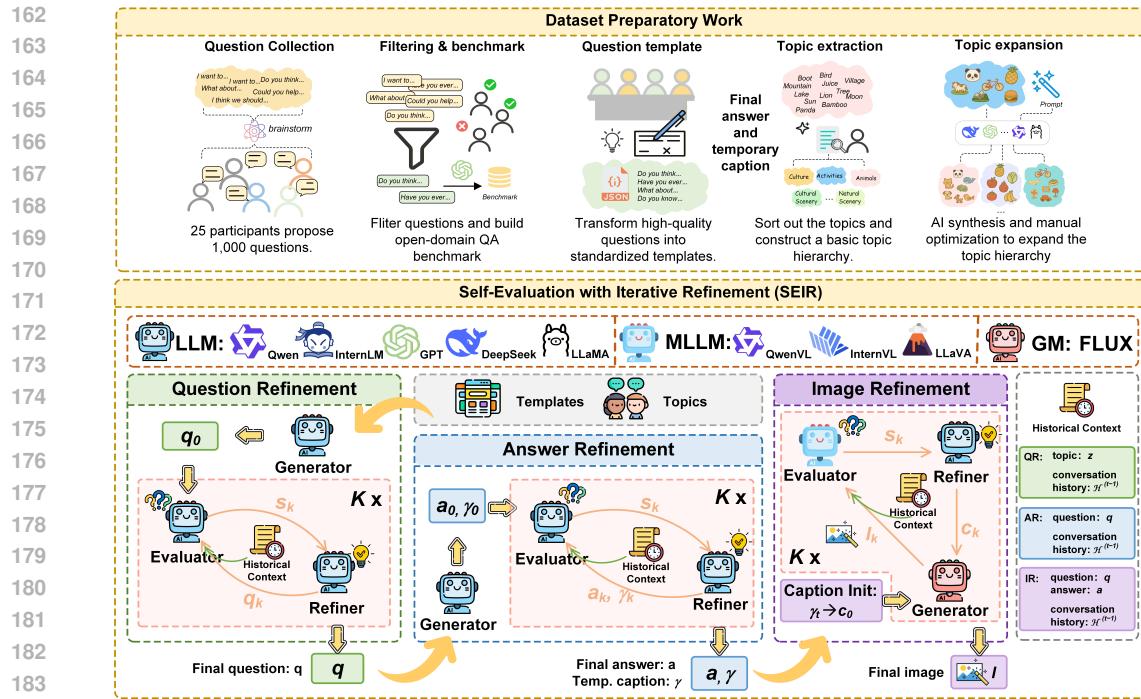


Figure 2: **Overview of the InterSyn Dataset Construction Framework.** The top panel illustrates the dataset preparatory work, covering question collection, filtering, template standardization, and topic expansion. The bottom panel illustrates the **Self-Evaluation with Iterative Refinement (SEIR)** method, which employs a *Generate-Evaluate-Refine* loop across three cascaded stages. (1) Question Refinement (QR): An initial question  $q_0$  is refined into the final question  $q$  based on the topic  $z$  and question template. (2) Answer Refinement (AR): Using  $q$ , the final answer  $a$  and a temporary caption  $\gamma$  are iteratively refined. (3) Image Refinement (IR): Initialize  $c_0$  with  $\gamma$ . Refine the caption and image until the final image  $I$  is produced. The right-side legend details the inputs and historical context ( $H^{(t-1)}$ ) used at each stage.

**Question Filtering and Benchmark.** We combined LLM-based filtering and expert review to select high quality questions. Redundant, ambiguous, uncommon, and overly subjective samples were removed based on predefined criteria. In total, 500 questions were reserved as a **basic** benchmark.

**Question Template Extraction.** Based on the selected high quality questions, we constructed a set of generalized question templates that capture the conversational query style (i.e., the common linguistic forms people use to pose requests), independent of any specific knowledge content, thereby enabling scalable question generation. See the appendix §E for details.

**Basic Topic Hierarchy.** We performed AI-assisted topic extraction from the filtered questions and manually organized the results to build a basic topic hierarchy, ensuring clear logical dependencies and coherent topic relations.

**Topic Hierarchy Expansion.** We further refined and expanded the basic topic hierarchy to improve both coverage and granularity. Combining AI-assisted topic suggestions with expert curation, we constructed a well-structured hierarchy that supports diverse and scalable data generation.

### 3.3 SEIR METHOD

The Self-Evaluation with Iterative Refinement (SEIR) method enhances data quality via a *Generate-Evaluate-Refine* loop. For each conversation turn  $t$ , we define a generic refinement operator  $\Phi$  that transforms an **initial content** (e.g., a question or answer)  $x_0$  into a **final content**  $x$  through  $K$  iterations. The loop terminates when the evaluator returns no suggestions or when a maximum iteration depth  $K$  is reached. The update rule at iteration  $k$  is:

216

217 
$$x_k = \mathcal{M}_{refine}(x_{k-1}, s_k), \quad \text{where } s_k = \mathcal{M}_{eval}(x_{k-1}, \mathcal{C}). \quad (1)$$
 218

219 Here,  $\mathcal{M}_{eval}$  acts as a judge, analyzing the current content  $x_{k-1}$  against context  $\mathcal{C}$  (e.g., the target  
220 topic, preceding question, or dialogue history) to provide specific suggestions  $s_k$ .  $\mathcal{M}_{refine}$  then  
221 produces an improved version  $x_k$ . We apply this operator sequentially across three stages for the  
222 current turn  $t$ :223 **Step 1: Question Refinement (QR).** We first generate an **initial question**  $q_0^{(t)}$  from a template and  
224 topic. To ensure clarity and depth, we apply  $\Phi$  to obtain the **final question**  $q^{(t)}$ :  
225

226 
$$q^{(t)} = \Phi(q_0^{(t)} \mid \mathcal{C} = \{z, \mathcal{H}^{(t-1)}\}). \quad (2)$$
 227

228 The evaluator critiques the initial question  $q_0^{(t)}$  based on the topic  $z$  and conversation history  $\mathcal{H}^{(t-1)}$ ,  
229 continuing the refinement until the question meets quality standards or the iteration limit is reached.  
230231 **Step 2: Answer Refinement (AR).** Using the final question  $q^{(t)}$ , we generate an **initial answer-**  
232 **caption pair**  $(a_0^{(t)}, \gamma_0^{(t)})$ . The refinement loop ensures the text is comprehensive and the caption is  
233 relevant, yielding the **final answer**  $a^{(t)}$  and a **temporary caption**  $\gamma^{(t)}$ :  
234

235 
$$(a^{(t)}, \gamma^{(t)}) = \Phi((a_0^{(t)}, \gamma_0^{(t)}) \mid \mathcal{C} = \{q^{(t)}, \mathcal{H}^{(t-1)}\}). \quad (3)$$
 236

237 **Step 3: Image Refinement (IR).** Finally, we generate and refine the image. We utilize the temporary  
238 caption  $\gamma^{(t)}$  from AR as the **initial caption**  $c_0^{(t)}$ . In each iteration  $k$ , we generate an image  $I_k^{(t)}$ , which  
239 is then critiqued by a VLM ( $\mathcal{V}$ ) against the textual context  $(q^{(t)}, a^{(t)})$  and dialogue history  $\mathcal{H}^{(t-1)}$ .  
240 The VLM’s feedback  $s_c^{(k)}$  guides the generation of a **refined caption**  $c_{k+1}^{(t)}$ :  
241

242 
$$I_k^{(t)} = \mathcal{G}(c_k^{(t)}), \quad s_c^{(k)} = \mathcal{V}(I_k^{(t)}, q^{(t)}, a^{(t)}, \mathcal{H}^{(t-1)}), \quad c_{k+1}^{(t)} = \mathcal{M}_{refine}(c_k^{(t)}, s_c^{(k)}). \quad (4)$$
 243

244 This cycle repeats until satisfaction or the iteration limit  $K$  is met, yielding the **final image**  $I^{(t)}$ .  
245

## 3.4 INTERSYN COMPOSITION

247 InterSyn contains approximately 1.8 million single-turn samples and 50k multi-turn dialogues. Data  
248 quality is ensured through the SEIR method, which iteratively refines samples to improve answer  
249 accuracy, conversational coherence, and image-text synergy. InterSyn offers a rare combination of  
250 diversity and quality, providing a robust foundation for training unified multimodal models with  
251 strong instruction-following and contextual reasoning capabilities.  
252

## 3.5 SYNJUDGE: A RELIABLE EVALUATOR FOR INTERLEAVED OUTPUTS

253 To reliably evaluate the complex, instruction-following capabilities of interleaved image-text gen-  
254 erators, we propose **SynJudge**, a reliable and comprehensive judge model that demonstrates high  
255 alignment with human judgment.  
256

## 3.5.1 EVALUATION DIMENSIONS

259 SynJudge is designed to provide interpretable, quantitative feedback by assessing responses across  
260 four key, complementary dimensions (See the appendix §F for details.):  
261262 **Text Content Completeness (TCC):** Assesses whether the generated text completely and accu-  
263 rately answers the user’s question, covering all sub-tasks and constraints.  
264265 **Image Content Completeness (ICC):** Assesses whether the generated images correctly depict the  
266 key objects, scenes, or concepts requested in the prompt.  
267268 **Image Quality (IQ):** Evaluates the visual fidelity, aesthetics, and overall quality of the generated  
269 images, penalizing artifacts, distortions, or low resolution.  
270271 **Image–Text Synergy (ITS):** Measures the cross-modal relationship. Unlike simple consistency  
272 metrics, ITS specifically rewards tight, complementary alignment where the text and images work  
273 together.  
274

270 together to form a cohesive answer. It penalizes redundancy (e.g., text merely describing the image)  
 271 and irrelevance.  
 272

273 **3.5.2 TRAINING AND MODEL SELECTION**  
 274

275 To ensure robust evaluation, we first constructed a high-quality, human-annotated dataset. The pro-  
 276 cess began by collecting a diverse set of interleaved responses from various generators (detailed in  
 277 §4.1.2). These responses were then rigorously scored by a panel of ten expert annotators across  
 278 the four dimensions (TCC, ICC, IQ, ITS). To ensure high reliability, each sample was scored by  
 279 multiple experts. Any scoring inconsistencies were resolved through a mandatory discussion pro-  
 280 tocol until a unified, final score was reached for each sample. This effort yielded a total of 48,000  
 281 annotated pairs, which were split into a **38,400-sample training set** and a held-out **9,600-sample**  
 282 **validation set**. Using this data, we fine-tuned strong MLLM backbones, including QwenVL and In-  
 283 ternalVL. As demonstrated in our experiments (see §4.4), the model trained from **QwenVL** achieved  
 284 the strongest alignment with human judgment, exhibiting the lowest RMSE and highest agreement  
 285 (A@1). Consequently, we designate this **QwenVL-trained** model as our final **SynJudge**.  
 286

287 **3.5.3 EVALUATION BENCHMARK FOR GENERATORS**  
 288

289 To rigorously assess the quality of interleaved image-text outputs, we established a fixed **Evaluation**  
 290 **Benchmark**. This benchmark comprises a curated series of **4,000** questions (500 human-authored  
 291 from §3.2 and 3,500 SEIR-generated) spanning the full topic hierarchy. By requiring different mod-  
 292 els to generate responses to this identical set of questions, we ensure a fair, standardized, and com-  
 293 prehensive comparison of their instruction-following and multimodal generation capabilities.  
 294

295 **4 EXPERIMENT**  
 296

297 **4.1 EXPERIMENTAL SETUP**  
 298

299 **4.1.1 EVALUATION DIMENSIONS FOR INTERLEAVED OUTPUTS**  
 300

301 We adopt the four evaluation dimensions defined in §3.5.1: TCC, ICC, IQ, and ITS. Briefly, TCC  
 302 and ICC assess content correctness in text and image respectively; IQ focuses on visual fidelity; and  
 303 ITS specifically measures the cross-modal synergy and alignment.  
 304

305 To quantitatively assess generators’ performance, we adopt the **mean score** (average score across  
 306 every dimension) and **variance score** (stability across questions). See the appendix §D.2 for details.  
 307

308 **4.1.2 EVALUATED GENERATORS**  
 309

310 We evaluate 13 multimodal generators  $G$  capable of producing interleaved image-text outputs. we  
 311 categorize them into two groups based on whether they natively support interleaved generation:  
 312

313 (1) **Non-Interleaved Generators:** These generators produce text and images sequentially through  
 314 modular pipelines, which include: Emu3 (Wang et al., 2024b), Janus-Pro (Chen et al., 2025b),  
 315 VILA-U (Wu et al., 2024), Show-o (Xie et al., 2024), Show-o-Turbo (Xu et al., 2025), Liquid (Wu  
 316 et al., 2025), D-DiT (Li et al., 2025), GPT-4o (OpenAI, 2024) + DALL-E3 (Betker et al., 2023), and  
 317 Gemini2.5 (Comanici et al., 2025) + FLUX (Black Forest Labs, 2024).  
 318

319 (2) **Interleaved Generators:** These generators can generate interleaved image-text outputs within a  
 320 unified process, including VARGPT (Zhuang et al., 2025a), VARGPT-v1.1 (Zhuang et al., 2025b),  
 321 Anole (Chern et al., 2024) and BAGEL (Deng et al., 2025).  
 322

323 **4.2 EFFICIENCY OF INTERSYN**  
 324

325 **4.2.1 DATA EFFICIENCY AND SCALABILITY**  
 326

327 We fine-tune four generators on randomly sampled InterSyn subsets of sizes 25k/50k/100k/200k and  
 328 evaluate with SynJudge. Table 1 shows consistent improvements as data increases. Notably, even  
 329 **25k/50k** samples already yield clear gains across all dimensions, and further scaling to **200k** contin-  
 330 ues to improve TCC, ICC, and especially ITS, which highlights the InterSyn dataset’s effectiveness  
 331 in enhancing both semantic alignment and answer completeness.  
 332

324  
 325 Table 1: Fine-tuning results on varying subset sizes of InterSyn. Performance consistently improves  
 326 as training data scales from 25K to 200K samples, demonstrating the dataset’s effectiveness and  
 327 scalability. Notably, just 50K samples yield substantial gains across all models, with continued im-  
 328 provement in content and synergy metrics (TCC, ICC, ITS) at larger scales. All scores are SynJudge  
 329 means.

Subset	Anole				VILA-U				VARGPT-v1.1				BAGEL			
	TCC	ICC	IQ	ITS	TCC	ICC	IQ	ITS	TCC	ICC	IQ	ITS	TCC	ICC	IQ	ITS
baseline	3.09	3.01	2.92	2.26	2.46	3.72	3.37	2.19	3.26	1.01	1.23	0.68	3.11	3.89	4.23	2.87
+ 25k	3.35	3.25	3.01	2.40	2.95	3.78	3.38	2.90	3.51	2.45	2.90	2.55	3.45	4.02	4.21	3.28
+ 50k	3.47	3.28	3.10	2.74	3.19	3.83	3.39	3.20	3.68	3.12	3.67	3.00	3.69	4.11	4.19	3.56
+ 100k	3.51	3.41	3.13	2.87	3.33	3.88	3.31	3.28	3.73	3.22	3.66	3.20	3.87	4.09	4.31	3.78
+ 200k	3.64	3.52	3.08	3.11	3.49	4.01	3.40	3.47	3.86	3.39	3.72	3.53	4.13	4.18	4.25	4.02

#### 335 336 4.2.2 RETENTION OF GENERAL UNDERSTANDING

337 Table 2: Understanding performance after  
 339 50k InterSyn fine-tuning. Values in paren-  
 340 theses denote the change ( $\Delta$ ) from the base.  
 341

Model	MME-P	MMBench	MMMU	SEEDBench
VILA-U	1344 (+8)	—	—	57.1 (+0.8)
VARGPT	1465 (-23)	66.8 (-0.8)	37.2 (+0.8)	65.6 (-2.3)
VARGPT-v1.1	1658 (-26)	79.4 (-1.6)	46.2 (-2.3)	75.2 (-0.9)
BAGEL	1646 (-41)	83.1 (-1.9)	52.8 (-2.5)	—

Crucially, the substantial gains in interleaved generation capabilities (demonstrated in §4.2.1) do not come at the cost of core understanding performance. As shown in Table 2, the models’ performance on standard understanding benchmarks (Fu et al., 2023; Liu et al., 2025; Yue et al., 2024; Li et al., 2024) remains robust after fine-tuning.

#### 345 346 4.2.3 EFFECTIVENESS OF MULTI-TURN DATA

347 To verify the effectiveness of our multi-turn dialogue data, we designed an experiment to assess its  
 348 impact on a model’s conversational capabilities. We fine-tuned two models capable of multi-turn  
 349 generation, Anole and VARGPT-v1.1, on different compositions of single-turn and multi-turn data,  
 350 while keeping the total training size fixed at 50k samples. The goal is to demonstrate that training  
 351 with multi-turn data enhances a model’s ability to maintain context and quality across an extended  
 352 conversation. The performance, evaluated by TCC, ICC, IQ, and ITS, is presented in Table 3.

354 Table 3: Effectiveness of multi-turn data on conversational performance across across different dia-  
 355 logue test turns. Models are trained on different proportions of single-turn and multi-turn data.

Model	Setting	Training Data		Turn 1			Turn 2			Turn 3					
		Single	Multi	TCC	ICC	IQ	ITS	TCC	ICC	IQ	ITS	TCC	ICC	IQ	ITS
Anole	Baseline	-	-	3.09	3.01	2.92	2.26	2.80	2.75	2.60	1.90	2.40	2.30	2.10	1.40
	Trained	50k	0	3.47	3.28	3.10	2.74	3.25	2.85	2.70	2.40	2.85	2.60	2.45	2.05
	Trained	25k	25k	3.48	3.27	3.13	2.77	3.35	3.00	2.80	2.40	3.00	2.70	2.55	2.10
	Trained	0	50k	3.52	3.24	3.10	2.94	3.40	3.05	2.90	2.55	3.27	2.85	2.70	2.25
VARGPT-v1.1	Baseline	-	-	3.26	1.01	1.23	0.68	3.10	0.95	1.18	0.72	2.90	0.97	0.90	0.65
	Trained	50k	0	3.68	3.12	3.67	3.00	3.40	2.90	3.45	2.90	3.05	2.60	3.05	2.65
	Trained	25k	25k	3.65	3.18	3.66	3.18	3.45	2.95	3.40	2.90	3.21	2.78	3.24	2.66
	Trained	0	50k	3.64	3.20	3.68	3.11	3.58	3.10	3.52	3.05	3.48	2.95	3.45	2.90

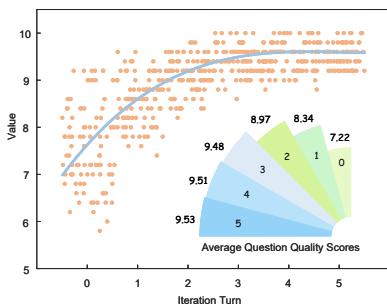
365 The results clearly demonstrate the value of multi-turn training data. First, the inclusion of multi-  
 366 turn data does not compromise first-turn performance, which remains high across all trained settings.  
 367 This is expected, as the SEIR pipeline ensures comparable data quality regardless of dialogue type.  
 368 Second, and more importantly, multi-turn training significantly reduces performance degradation in  
 369 later turns. Models trained only on single-turn data show a steeper drop in quality as the conversa-  
 370 tion progresses. In contrast, training with multi-turn dialogues mitigates this degradation, particu-  
 371 larly for ITS, by explicitly teaching the model to maintain context across extended conversational  
 372 dependencies. This confirms the effectiveness of our multi-turn dataset in fostering more coherent  
 373 and consistent multi-turn generation.

#### 374 375 4.3 EFFECTIVENESS OF SEIR

376 To validate the effectiveness of the SEIR, we conduct experiments on both the iterative refinement  
 377 process and the final output quality comparison with other generators.

378 4.3.1 VALIDATION OF ITERATION REFINEMENT  
379

380 In this section, all evaluations in this section are conducted by human judge to establish a ground  
381 truth. For question quality, (Figure 3) shows that QR improves quality over the first three iterations  
382 but plateaus thereafter. Based on this, we set the QR to 3, achieving 99.5% of peak quality while  
383 reducing computational cost by 40%. For answer quality, we evaluate across different iterations  
384 of AR and IR, using a set of 7,000 questions generated through three rounds of QR. As shown in  
385 Table 4, the results confirm that AR primarily improves content completeness, while IR enhances  
386 multimodal synergy, demonstrating the effectiveness of the SEIR method. Experimental results  
387 show that when both AR and IR are set to 4 or 5, the improvements become marginal. Based on  
388 this, we fix the number of AR and IR iterations to 3 in dataset construction settings.



390  
391  
392  
393  
394  
395  
396  
397  
398  
399  
400  
401  
402  
403  
404  
405  
406  
407  
408  
409  
410  
411  
412  
413  
414  
415  
416  
417  
418  
419  
420  
421  
422  
423  
424  
425  
426  
427  
428  
429  
430  
431  
Figure 3: **Impact of question refinement (QR) on question quality.** This plot shows the quality scores across different QR iterations. Quality improves significantly over the first three iterations but plateaus thereafter.

Table 4: Impact of answer refinement (AR) and image refinement (IR) on answer quality. The table reports human evaluation **mean** scores across four dimensions (TCC, ICC, IQ, ITS). AR improves TCC, ICC, and ITS, while IR further enhances ICC and ITS, confirming the effectiveness of iterative refinement.

AR	IR	TCC	ICC	IQ	ITS
0	0	3.85	4.01	4.42	3.79
1	1	4.19	4.21	4.38	4.07
2	2	4.34	4.35	4.41	4.38
3	0	4.42	4.11	4.37	4.04
3	1	4.42	4.33	4.43	4.35
3	2	4.42	4.43	4.39	4.46
3	3	4.42	4.47	4.44	4.52
4	4	4.44	4.50	4.45	4.53
5	5	4.45	4.49	4.43	4.53

## 4.3.2 COMPARISON OF SEIR OUTPUTS WITH OTHER GENERATORS

420 We evaluate the SEIR pipeline against 13 baseline generators using the fixed **Evaluation Benchmark**  
421 defined in §3.5.3. In this section, we employ both human judges and SynJudge to assess the  
422 generator outputs across the four evaluation dimensions.

423 As shown in Table 5, both human judges and SynJudge confirm that SEIR-generated samples (Inter-  
424 Syn) achieve the highest mean scores across all dimensions. They outperform the strongest baseline,  
425 GPT-4o+DALL-E, by a margin of 0.34–0.66, with the largest gap observed in ITS. Furthermore,  
426 InterSyn exhibits very low variance (below 0.61), reflecting its consistent output quality and the  
427 robustness of our automated method. Crucially, the significant performance gap between InterSyn  
428 and the best SOTA generators reveals that even top models still struggle with image–text alignment  
429 and complementarity, indicating substantial room for future improvement.

## 4.4 RELIABILITY OF SYNJUDGE

## 4.4.1 EVALUATION SETUP FOR JUDGES

430 To identify a reliable automatic evaluator that aligns closely with human scoring, we conduct a  
431 comparative experiment. The protocol is designed to rigorously measure each candidate judge’s  
432 deviation from a human-annotated ground truth.

433 We evaluate a total of five model-based judges against our human ground truth. The candidates  
434 are: (1) **Human Judge (Ground Truth):** A panel of ten experts whose scores serve as the gold  
435 standard. A cross-review protocol was used to ensure scoring reliability and mitigate individual  
436 bias. (2) **Zero-Shot MLLM Judges:** Three off-the-shelf MLLMs used for automated assessment:  
437 GPT-4o (OpenAI, 2024), QwenVL2.5 (Bai et al., 2025), and InternVL2.5 (Chen et al., 2024c).  
438 (3) **SynJudge Candidates (Finetuned):** We fine-tuned two strong MLLM backbone candidates,  
439 **QwenVL-trained** and **InternVL-trained**, to create our proposed evaluator.

440 The evaluation is conducted on a test set of 9,600 human-annotated question-answer pairs. To  
441 provide a comprehensive assessment of judge performance, we use two complementary metrics: (1)  
442 Root Mean Squared Error (RMSE): Measures the magnitude of the deviation from human scores. (2)

432 Table 5: Generator performance evaluated by human judge and SynJudge. Each entry is reported  
 433 as mean (variance), where the value outside the parentheses denotes the mean score and the value  
 434 inside the parentheses denotes the variance.

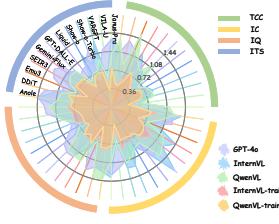
Generator	Human				SynJudge			
	TCC	ICC	IQ	ITS	TCC	ICC	IQ	ITS
Anole	3.06 (1.47)	2.95 (2.05)	2.89 (1.87)	2.25 (2.47)	3.09 (1.47)	3.01 (2.11)	2.92 (1.93)	2.26 (2.55)
DDiT	0.38 (1.17)	3.51 (0.83)	3.29 (0.85)	0.37 (1.21)	0.28 (0.86)	3.67 (0.75)	3.34 (0.92)	0.26 (0.87)
Emu3	3.38 (0.86)	3.86 (0.86)	3.87 (0.59)	3.26 (1.31)	3.37 (0.74)	3.92 (0.66)	3.85 (0.80)	3.37 (1.16)
VILA-U	2.45 (2.58)	3.62 (0.91)	3.35 (0.97)	2.31 (3.01)	2.46 (2.73)	3.72 (0.92)	3.37 (1.12)	2.19 (2.96)
Liquid	2.88 (0.76)	3.82 (0.69)	3.67 (0.70)	3.00 (1.42)	2.87 (0.71)	3.86 (0.75)	3.76 (0.85)	3.06 (1.36)
Janus-Pro	2.93 (0.95)	3.25 (1.21)	3.16 (0.99)	2.55 (1.75)	2.96 (0.96)	3.30 (1.09)	3.11 (1.11)	2.62 (1.87)
Show-o	3.32 (1.31)	3.65 (0.94)	3.52 (0.89)	3.10 (1.94)	3.49 (1.12)	3.79 (0.82)	3.57 (0.96)	3.30 (1.70)
Show-o-Turbo	3.48 (1.12)	3.77 (0.88)	3.53 (0.94)	3.33 (1.60)	3.50 (1.06)	3.89 (0.86)	3.64 (1.01)	3.45 (1.44)
VARGPT	2.60 (0.66)	0.94 (2.37)	0.94 (2.32)	0.55 (1.99)	2.55 ( <b>0.30</b> )	0.94 (2.39)	0.87 (2.17)	0.67 (2.16)
VARGPT-v1.1	3.13(0.83)	0.89(1.98)	1.16(2.11)	0.72(1.83)	3.26(0.76)	1.01(2.12)	1.23(1.95)	0.68(1.96)
BAGEL	2.97(0.83)	3.92(0.81)	4.18(0.75)	2.81(1.21)	3.11(0.91)	3.89(0.72)	4.23(0.66)	2.87(1.33)
Gemini+Flux	3.94 (0.94)	4.06 (0.58)	4.43 ( <b>0.47</b> )	3.81 (0.90)	3.97 (0.57)	4.12 (0.71)	<b>4.48</b> (0.64)	3.84 (1.11)
GPT-4o+DALL-E	4.05 ( <b>0.37</b> )	4.08 ( <b>0.48</b> )	4.41 (0.57)	3.94 (0.64)	3.99 (0.65)	4.10 (0.81)	4.45 (0.58)	3.87 (1.16)
SEIR	<b>4.41</b> (0.55)	<b>4.46</b> (0.55)	<b>4.47</b> (0.53)	<b>4.51</b> ( <b>0.57</b> )	<b>4.39</b> (0.61)	<b>4.49</b> (0.63)	<b>4.44</b> ( <b>0.45</b> )	<b>4.53</b> ( <b>0.51</b> )

451  
 452 Human Agreement (A@1): Measures the percentage of scores that are within a 1-point tolerance of  
 453 the human rating, reflecting practical reliability. A lower RMSE and a higher A@1 indicate stronger  
 454 alignment with human judgment. The SynJudge candidates were trained on a separate training set  
 455 of 38,400 human-annotated pairs. Detailed definitions of the metrics and training hyperparameters  
 456 of SynJudge are provided in the appendix (§D.2, §D.4).

#### 458 4.4.2 COMPARISON RESULTS

459 Figure 4 and Table 6 report the performance of each judge using both RMSE and A@1 metrics. The  
 460 results show a clear trend: finetuned judges decisively outperform zero-shot models on both metrics.

462 The **QwenVL-trained** judge demonstrates the strongest alignment with human preferences, achieving  
 463 both the lowest average RMSE and the highest average A@1 of **95.4%**. This high agreement rate  
 464 signifies that over 95% of its scores are within one point of human judgment, confirming its high re-  
 465 liability. In contrast, zero-shot judges like GPT-4o lag significantly, with A@1 scores around 86.5%.  
 466 Based on this superior performance across complementary metrics, we select QwenVL-trained as  
 467 our final **SynJudge**.



477 Figure 4: Visualization of  
 478 RMSE for different judges.

479 Table 6: Judge performance comparison. We report average  
 480 RMSE (lower is better) and Human Agreement (A@1, higher  
 481 is better) against human scores. The best result in each row is  
 482 highlighted in bold. QwenVL-trained demonstrates the strongest  
 483 alignment.

Dimension	GPT-4o	QwenVL	InternVL	QwenVL-trained	InternVL-trained
TCC (RMSE)	1.01	0.81	0.96	<b>0.54</b>	0.55
ICC (RMSE)	1.02	1.09	0.90	0.72	<b>0.70</b>
IQ (RMSE)	0.98	0.96	1.06	<b>0.68</b>	0.72
ITS (RMSE)	1.18	1.20	1.03	<b>0.67</b>	0.72
A@1	0.865	0.875	0.866	<b>0.954</b>	0.945

## 484 5 CONCLUSION

485 In this paper, we present InterSyn, a large scale, high quality multimodal dataset designed for  
 486 instruction-following and interleaved image-text generation. Constructed via the fully automated  
 487 SEIR method, InterSyn combines scale, diversity, and fidelity, supporting multi-turn dialogues  
 488 where each response is refined to achieve not only semantic completeness but also tight image-text  
 489 synergy—ensuring that visual and textual modalities complement each other to convey meaning col-

486 laboratively. To complement InterSyn, we introduce SynJudge, a multi-dimensional automatic eval-  
487 uator specifically designed to assess interleaved outputs across four key dimensions, including a ded-  
488 icated metric for image-text synergy. Unlike traditional metrics focused on surface-level alignment  
489 or consistency, SynJudge emphasizes the semantic interplay between images and text, rewarding  
490 complementary relationships while penalizing redundancy or disjointness. Extensive experiments  
491 validate the effectiveness of both InterSyn and SEIR. Models fine-tuned on InterSyn consistently  
492 outperform strong baselines, showing notable improvements in instruction alignment, multimodal  
493 reasoning, and especially the ability to produce coherent, synergistic interleaved content. We be-  
494 lieve this work lays a solid foundation for future research in scalable multimodal data generation,  
495 robust synergy-centric evaluation, and the development of general-purpose multimodal intelligence  
496 systems that understand and communicate across modalities in a truly integrated manner.  
497  
498  
499  
500  
501  
502  
503  
504  
505  
506  
507  
508  
509  
510  
511  
512  
513  
514  
515  
516  
517  
518  
519  
520  
521  
522  
523  
524  
525  
526  
527  
528  
529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539

540     **Reproducibility Statement.** To facilitate the verification of our findings and to support future re-  
 541     search, we are committed to making our work fully reproducible. The complete codebase, including  
 542     scripts for the SEIR data generation pipeline, SynJudge training, and all evaluation protocols, will be  
 543     made publicly available on GitHub upon publication. We will also release the full InterSyn dataset  
 544     (1.8M samples), all benchmark sets used for our main experiments, and the trained weights of our  
 545     final SynJudge evaluator. The core of our methodology relies on publicly available models, and  
 546     all prompts, model configurations, and hyperparameters are extensively detailed in the appendix to  
 547     ensure that our results can be precisely replicated.

548     

549     **REFERENCES**

550     

551     Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel  
 552     Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language  
 553     model for few-shot learning. In *Proceedings of the Advances in Neural Information Processing  
 554     Systems*, 2022.

555     Jie An, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Lijuan Wang, and Jiebo  
 556     Luo. Openleaf: A novel benchmark for open-domain interleaved image-text generation. In *Pro-  
 557     ceedings of the 32nd ACM International Conference on Multimedia*, MM '24, pp. 11137–11145,  
 558     New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706868. doi:  
 559     10.1145/3664647.3685511. URL <https://doi.org/10.1145/3664647.3685511>.

560     Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang,  
 561     Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan,  
 562     Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng,  
 563     Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv  
 564     preprint arXiv:2502.13923*, 2025.

565     James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang  
 566     Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer  
 567     Science*. <https://cdn.openai.com/papers/dall-e-3.pdf>, 2(3):8, 2023.

568     Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024. Ac-  
 569     cessed: 2024-11-05.

570     Dongping Chen, Ruoxi Chen, Shu Pu, Zhaoyi Liu, Yanru Wu, Caixi Chen, Benlin Liu, Yue Huang,  
 571     Yao Wan, Pan Zhou, and Ranjay Krishna. Interleaved scene graphs for interleaved text-and-image  
 572     generation assessment, 2025a. URL <https://arxiv.org/abs/2411.17188>.

573     Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahu-  
 574     Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Confer-  
 575     ence on Computer Vision*, pp. 370–387. Springer, 2024a.

576     Wei Chen, Lin Li, Yongqi Yang, Bin Wen, Fan Yang, Tingting Gao, Yu Wu, and Long Chen. CoMM:  
 577     A coherent interleaved image-text dataset for multimodal understanding and generation. *arXiv  
 578     preprint arXiv:2406.10462*, 2024b.

579     Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and  
 580     Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model  
 581     scaling. *arXiv preprint arXiv:2501.17811*, 2025b.

582     Yixiong Chen, Li Liu, and Chris Ding. X-iqe: explainable image quality evaluation for text-to-  
 583     image generation with visual large language models, 2023. URL <https://arxiv.org/abs/2305.10843>.

584     Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shen-  
 585     glong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source  
 586     multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*,  
 587     2024c.

594 Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong  
 595 Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning  
 596 for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer  
 597 Vision and Pattern Recognition*, pp. 24185–24198, 2024d.

598 Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. Anole: An open, autoregressive, native large multi-  
 599 modal models for interleaved image-text generation. *arXiv preprint arXiv:2407.06135*, 2024.

600 Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, and so on. Gemini 2.5: Pushing the frontier  
 601 with advanced reasoning, multimodality, long context, and next generation agentic capabilities,  
 602 2025. URL <https://arxiv.org/abs/2507.06261>.

603 Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Wei-  
 604 hao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified  
 605 multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

606 Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam  
 607 Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for  
 608 high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*,  
 609 2024a.

610 Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam  
 611 Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for  
 612 high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*,  
 613 2024b.

614 Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu  
 615 Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal  
 616 large language models. *arXiv preprint arXiv:2306.13394*, 2023.

617 Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore:  
 618 A reference-free evaluation metric for image captioning. In Marie-Francine Moens, Xuan-  
 619 jing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Confer-  
 620 ence on Empirical Methods in Natural Language Processing*, pp. 7514–7528, Online and  
 621 Punta Cana, Dominican Republic, November 2021. Association for Computational Linguis-  
 622 tics. doi: 10.18653/v1/2021.emnlp-main.595. URL <https://aclanthology.org/2021.emnlp-main.595>.

623 Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter.  
 624 Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in  
 625 neural information processing systems*, 30, 2017.

626 Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob  
 627 Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In  
 628 *Proceedings of the 2016 conference of the North American chapter of the association for compu-  
 629 tational linguistics: Human language technologies*, pp. 1233–1239, 2016.

630 Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov,  
 631 Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open  
 632 web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information  
 633 Processing Systems*, 36:71683–71702, 2023.

634 Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan.  
 635 Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF  
 636 Conference on Computer Vision and Pattern Recognition*, pp. 13299–13308, 2024.

637 Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image  
 638 pre-training with frozen image encoders and large language models. In *International conference  
 639 on machine learning*, pp. 19730–19742. PMLR, 2023.

640 Zijie Li, Henry Li, Yichun Shi, Amir Barati Farimani, Yuval Kluger, Linjie Yang, and Peng Wang.  
 641 Dual diffusion for unified image generation and understanding, 2025. URL <https://arxiv.org/abs/2501.00289>.

648 Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization*  
 649 *branches out*, pp. 74–81, 2004.  
 650

651 Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. Revisiting the  
 652 role of language priors in vision-language models, 2024. URL <https://arxiv.org/abs/2306.01879>.  
 653

654 Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.  
 655

656 Minqian Liu, Zhiyang Xu, Zihao Lin, Trevor Ashby, Joy Rimchala, Jiaxin Zhang, and Lifu Huang.  
 657 Holistic evaluation for interleaved text-and-image generation. *arXiv preprint arXiv:2406.14643*,  
 658 2024.

659 Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan,  
 660 Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around  
 661 player? In *European Conference on Computer Vision*, pp. 216–233. Springer, 2025.

662 Yujie Lu, Xianjun Yang, Xiujun Li, Xin Eric Wang, and William Yang Wang. Llmscore: Unveiling  
 663 the power of large language models in text-to-image synthesis evaluation. *Advances in Neural*  
 664 *Information Processing Systems*, 36, 2024.

665 OpenAI. Hello GPT-4o. <https://openai.com/index/hello-gpt-4o/>, 2024. Accessed:  
 666 2024-05-26.

667 Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic  
 668 evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association*  
 669 *for Computational Linguistics*, pp. 311–318, 2002.

670 Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-  
 671 conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

672 Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and  
 673 Xi Chen. Improved techniques for training gans. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon,  
 674 and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Cur-  
 675 ran Associates, Inc., 2016. URL [https://proceedings.neurips.cc/paper\\_files/paper/2016/file/8a3363abe792db2d8761d6403605aeb7-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2016/file/8a3363abe792db2d8761d6403605aeb7-Paper.pdf).  
 676

677 Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint*  
 678 *arXiv:2405.09818*, 2024.

679 Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu,  
 680 Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng  
 681 Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model’s  
 682 perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

683 Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan  
 684 Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need.  
 685 *arXiv preprint arXiv:2409.18869*, 2024b.

686 Junfeng Wu, Yi Jiang, Chuofan Ma, Yuliang Liu, Hengshuang Zhao, Zehuan Yuan, Song Bai, and  
 687 Xiang Bai. Liquid: Language models are scalable and unified multi-modal generators, 2025.  
 688 URL <https://arxiv.org/abs/2412.04332>.  
 689

690 Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng  
 691 Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual  
 692 understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024.

693 Peng Xia, Siwei Han, Shi Qiu, Yiyang Zhou, Zhaoyang Wang, Wenhao Zheng, Zhaorun Chen,  
 694 Chenhang Cui, Mingyu Ding, Linjie Li, et al. Mmie: Massive multimodal interleaved compre-  
 695 hension benchmark for large vision-language models. *arXiv preprint arXiv:2410.10139*, 2024.

696 Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin,  
 697 Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer  
 698 to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.

702 Chenkai Xu, Xu Wang, Zhenyi Liao, Yishun Li, Tianqi Hou, and Zhijie Deng. Show-o turbo:  
 703 Towards accelerated unified multimodal understanding and generation, 2025. URL <https://arxiv.org/abs/2502.05415>.  
 704

705 Zhiyang Xu, Minqian Liu, Ying Shen, Joy Rimchala, Jiaxin Zhang, Qifan Wang, Yu Cheng, and  
 706 Lifu Huang. Lateralization LoRA: Interleaved instruction tuning with modality-specialized adap-  
 707 tations. *arXiv preprint arXiv:2407.03604*, 2024.  
 708

709 Yue Yang, Artemis Panagopoulou, Qing Lyu, Li Zhang, Mark Yatskar, and Chris Callison-Burch.  
 710 Visual goal-step inference using wikiHow. In Marie-Francine Moens, Xuanjing Huang, Lucia  
 711 Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods  
 712 in Natural Language Processing*, pp. 2167–2179, Online and Punta Cana, Dominican Republic,  
 713 November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.  
 714 165. URL <https://aclanthology.org/2021.emnlp-main.165>.  
 715

716 Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens,  
 717 Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun,  
 718 Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and  
 719 Wenhui Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning  
 720 benchmark for expert agi. In *Proceedings of CVPR*, 2024.  
 721

722 Kaizhi Zheng, Xuehai He, and Xin Eric Wang. MiniGPT-5: Interleaved vision-and-language gener-  
 723 ation via generative vokens. *arXiv preprint arXiv:2310.02239*, 2023.  
 724

725 Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from  
 726 web instructional videos. In Sheila A. McIlraith and Kilian Q. Weinberger (eds.), *Proceedings  
 727 of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), the 30th innovative  
 728 Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational  
 729 Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*,  
 730 pp. 7590–7598. AAAI Press, 2018. doi: 10.1609/AAAI.V32I1.12342. URL <https://doi.org/10.1609/aaai.v32i1.12342>.  
 731

732 Pengfei Zhou, Xiaopeng Peng, Jiajun Song, Chuanhao Li, Zhaopan Xu, Yue Yang, Ziyao Guo,  
 733 Hao Zhang, Yuqi Lin, Yefei He, Lirui Zhao, Shuo Liu, Tianhua Li, Yuxuan Xie, Xiaojun Chang,  
 734 Yu Qiao, Wenqi Shao, , and Kaipeng Zhang. Gate opening: A comprehensive benchmark for  
 735 judging open-ended interleaved image-text generation, 2024.  
 736

737 Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Young-  
 738 jae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open,  
 739 billion-scale corpus of images interleaved with text. In *Thirty-seventh Conference on Neu-  
 740 ral Information Processing Systems Datasets and Benchmarks Track*, 2023. URL <https://openreview.net/forum?id=t0d8rSjcWz>.  
 741

742 Xianwei Zhuang, Yuxin Xie, Yufan Deng, Liming Liang, Jinghan Ru, Yuguo Yin, and Yuexian  
 743 Zou. Vargpt: Unified understanding and generation in a visual autoregressive multimodal large  
 744 language model, 2025a. URL <https://arxiv.org/abs/2501.12327>.  
 745

746 Xianwei Zhuang, Yuxin Xie, Yufan Deng, Dongchao Yang, Liming Liang, Jinghan Ru, Yuguo  
 747 Yin, and Yuexian Zou. Vargpt-v1.1: Improve visual autoregressive large unified model via it-  
 748 erative instruction tuning and reinforcement learning, 2025b. URL <https://arxiv.org/abs/2504.02949>.  
 749

750

751

752

753

754

755

756 **A STATEMENT ON LLMs USAGE**  
757758 The authors used large language models (LLMs) during the writing process solely for language  
759 refinement and editing. It should be explicitly stated that LLMs were not employed in any core  
760 aspects of the research, including the formulation of research ideas, the design of methodologies,  
761 the execution of experiments, or the development of conclusions. All scholarly contributions were  
762 made independently by the authors.  
763764 **B APPENDIX OVERVIEW**  
765766 The all supplementary document is organized as follows:  
767768 

- 769 • Comparison of datasets and Samples of InterSyn are shown in §C.
- 770 • Supplementary analysis of experimental Data are shown in §D.
- 771 • Question templates and topic hierarchy are shown in §E.
- 772 • The evaluation dimensions for question and answer are shown in §F.
- 773 • Prompts used in this work are shown in §G.
- 774 • The benchmark samples are shown in §H.
- 775 • Human annotation platform are shown in §I.
- 776 • Limitations of this study are shown in §J.

  
778779 **C COMPARISON OF DATASETS AND INTERSYN SAMPLES**  
780781 **C.1 COMPARISON OF DATASETS**  
782783 The table provides a comprehensive comparison of InterSyn with representative multimodal datasets  
784 and benchmarks. Existing datasets such as MMC4 and OBELICS primarily rely on large scale web-  
785 crawled corpora, often lacking instruction-following capabilities and multi-turn structures. Other  
786 resources like CoMM and ShareGPT4V improve data cleanliness but remain limited to single-turn  
787 interactions without tight semantic supervision.  
788789 Recent efforts including LeafInstruct introduce instruction-following supervision but still operate  
790 in single-turn formats. Meanwhile, benchmark-oriented resources—such as OpenLEAF, ISG-  
791 BENCH, MMIE, InterleavedBench, and OpenING—focus on evaluating generation quality but are  
792 constrained by small scale and limited turn complexity.  
793794 In contrast, InterSyn is the first to offer a large scale, multi-turn, instruction-following dataset specif-  
795 ically designed for interleaved image-text generation. Built with the SEIR method, InterSyn not only  
796 ensures high quality visual-textual synergy but also scales to 1.8 million samples—orders of magni-  
797 tude larger than existing benchmarks. Its emphasis on dialogue coherence, iterative refinement, and  
798 synergistic multimodal responses fills a critical gap in current resources, laying the groundwork for  
799 developing and evaluating truly unified multimodal generation models.  
800801 Table 7: Comparison of Multimodal Datasets and Benchmarks. Abbreviations: Cat. = Category;  
802 Inst. = *instruction-following*; MT. = *multi-turn*; DS. = *dataset*; BM. = *benchmark*; Gen. = *gen-  
803 eration*.  
804805 

Name	Cat.	Source	Method	Size	Inst.	MT.
InterSyn	DS.	Collected questions	SEIR	1.8M	✓	✓
MMC4	DS.	Common Crawl	CLIP-based filtering	101.2M Doc.	✗	✗

	Name	Cat.	Source	Method	Size	Inst.	MT.
810	OBELICS	DS.	Common Crawl	Multi-granularity filtering	141M pages	$\times$	$\times$
811	CoMM	DS.	WikiHow, StoryBird, eHow, etc.	Multi-perspective filtering	227K Doc.	$\times$	$\times$
812	ShareGPT4V	DS.	GPT-4V captions	Share-Captioner	1.2M pairs	$\times$	$\times$
813	LeafInstruct	DS.	MMC4, VIST, YouCook2, etc.	Text & image quality filtering	38,272	$\checkmark$	$\times$
814	OpenLEAF	BM.	User Queries	GPT-4 Gen. and human review	660	$\checkmark$	$\times$
815	ISG-BENCH	BM.	VIST, CoMM, manual Gen.	Model Gen. & human review	1,150	$\checkmark$	$\times$
816	MMIE	BM.	WikiHow, VIST, Math-Vista, etc.	Sampling & reconstruction	20,103	$\checkmark$	$\times$
817	InterleavedBench	BM.	VIST, WikiHow, etc.	GPT-4o Gen. + human review	815	$\checkmark$	$\times$
818	OpenING	BM.	YouTube, Google, etc.	Manual pipeline	5,400	$\checkmark$	$\times$
819							
820							
821							
822							
823							
824							
825							
826							
827							
828							
829							
830							
831							
832							
833							
834							
835							
836							
837	C.2 SINGLE-TURN SAMPLES						
838	Samples of data are shown in Figure 5						
839							
840	C.3 MULTI-TURN SAMPLES						
841	Samples of data are shown in Figure 6						
842							
843							
844	D SUPPLEMENTARY ANALYSIS OF EXPERIMENTAL						
845							
846	D.1 SYMBOLS AND NOTATIONS FOR SEIR METHOD						
847							
848	We summarize the key symbols used in the SEIR method below:						
849							
850	• $\mathcal{T}$ : Set of question templates.						
851	• $\mathcal{Z}$ : Set of topics.						
852	• $T \in \mathbb{N}^+$ : Number of conversation turns.						
853	• $K \in \mathbb{N}^+$ : Number of refinement iterations at each stage.						
854	• $\mathcal{H}^{(t-1)}$ : History of the conversation up to turn $t-1$ , represented as $\{(q^{(i)}, a^{(i)}, i^{(i)})\}_{i=1}^{t-1}$ .						
855	• $q^{(t)}$ : <b>Final</b> question generated at conversation turn $t$ .						
856	• $q_k^{(t)}$ : Question after $k$ refinement iterations at conversation turn $t$ .						
857	• $a^{(t)}$ : <b>Final</b> text answer generated at conversation turn $t$ .						
858	• $a_k^{(t)}$ : Text answer after $k$ refinement iterations at conversation turn $t$ .						
859	• $\gamma^{(t)}$ : Temporary caption associated with the text answer at conversation turn $t$ .						
860	• $\gamma_k^{(t)}$ : Temporary caption after $k$ refinement iterations.						
861							
862							
863							

864  
865  
866  
867  
868  
869  
870  
871  
872  
873  
874

866  What fun sand game do kids enjoy at the beach? Any super fun ones you can describe or even show a picture of?

867  
868  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917

Building sandcastles is a beloved beach activity for kids, involving creative construction with molds or simple tools to form castles, complete with moats, towers, and decorations like shells and flags. Safety tips include using sun protection and keeping hydrated. Variations include themed castles or competitive castle-building games.



866  Could you describe the peaceful scene of a canyon under the stars for me? I'd love to picture it clearly in my mind.

867  
868  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917

A tranquil canyon under the starlit sky, where the scent of pine fills the air, the rustling of leaves whispers in the breeze, and the cool night air wraps around you, creating an immersive, serene experience.



866  Can you describe what it looks like when the sharp outlines of snow-covered peaks stand out against a crystal-clear night sky? It sounds breathtaking! Could you also show me an image of this scene?

867  
868  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917

Imagine standing beneath a star-studded, crystal-clear night sky, where the snow-topped mountains rise sharply against the deep blue heavens. The gentle moonlight bathes the peaks in a silvery sheen, their crisp outlines starkly defined, and the air is so clear it feels almost tangible, enveloping you in a serene calm.



866  What does a deer look like when it's covered in a light dusting of snow?

867  
868  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917

A deer with a light dusting of snow has a delicate, white layer that highlights its fur texture, creating a glistening effect that emphasizes its body contours and stark antlers, as if it's softly glowing against the winter backdrop.



866  Could you describe the scene of a stone house nestled among the mountains and surrounded by bright red leaves? And please create an image for me as well!

867  
868  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917

A charming stone house nestles cozily among towering mountains, embraced by a sea of vivid, bright red leaves, crafting a breathtakingly beautiful and serene autumnal scene.



866  Could you describe for me a scene where cherry blossom petals are floating on a tranquil pond with their reflections in the water? Also, please show it with an image.

867  
868  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917

On the glassy surface of the pond, delicate cherry blossom petals drift gently, their soft pink hues mirrored perfectly below, creating a dreamy, peaceful sight.

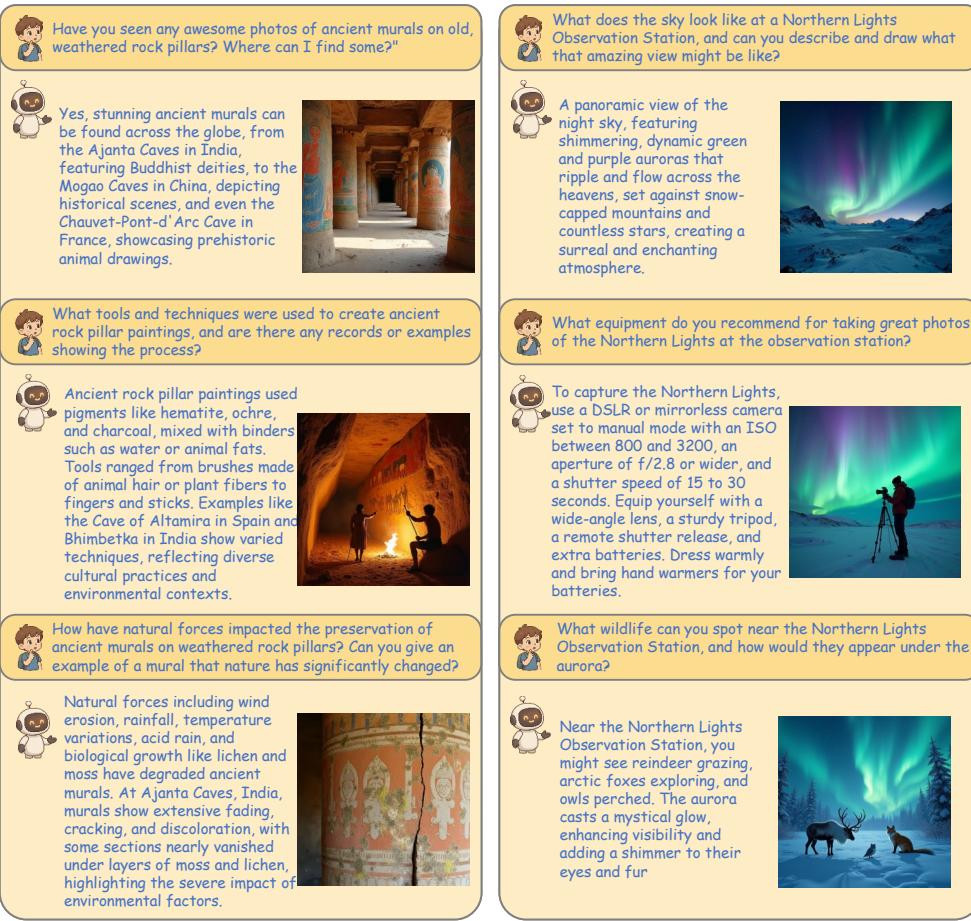


Figure 5: Examples of single-turn conversation

- $c^{(t)}$ : **Final** image caption at conversation turn  $t$ .
- $c_k^{(t)}$ : Image caption after  $k$  refinement iterations.
- $I^{(t)}$ : **Final** generated image at conversation turn  $t$ .
- $I_k^{(t)}$ : Generated image after  $k$  refinement iterations.
- $\mathcal{M}_L$ : Language model used for text generation and refinement.
- $\mathcal{M}_V$ : Vision-language model used for image caption evaluation and refinement.
- $\mathcal{M}_G$ : Text-to-image generation model.
- $p_g(\cdot)$ : Prompt function for generating model response.
- $p_s(\cdot)$ : Prompt function for generating refinement suggestions.
- $p_r(\cdot)$ : Prompt function for applying refinements.

## D.2 EVALUATION METRICS FOR JUDGES AND GENERATORS

To facilitate a quantitative evaluation of judges and generators, we design a set of metrics.

918  
919  
920  
921  
922  
923  
924  
925  
926  
927

928

929

930

931  
932  
933  
934  
935  
936  
937938  
939  
940941  
942  
943  
944  
945  
946  
947948  
949  
950  
951  
952

**Mean Calculation** Let  $x_{i,d}$  denote the score given by a judge to the  $i$ -th sample generated by a generator, under evaluation dimension  $d$ , and let  $N$  be the total number of samples. Then, for each (judge, generator) pair, the mean score is computed as:

$$S_d = \frac{1}{N} \sum_{i=1}^N x_{i,d} \quad (5)$$

961  
962

The mean score  $S_d$  reflects the average performance of a generator, as evaluated by a specific judge under dimension  $d$ .

963

**Variance Calculation** To estimate the variability of the scores, we compute the variance:

964  
965  
966  
967  
968  
969  
970

$$\sigma_d = \frac{1}{N} \sum_{i=1}^N (x_{i,d} - S_d)^2 \quad (6)$$

971

The variance  $\sigma_d$  captures the consistency of the generator's performance across different questions. A higher variance indicates greater inconsistency in quality.

972 **Root Mean Squared Error (RMSE)** To measure the agreement between a model-based judge  $M$   
 973 and human judge  $H$ , we compute the RMSE between their respective scores for each sample:  
 974

$$975 \quad 976 \quad 977 \quad 978 \quad \text{RMSE}_d = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_{i,d}^M - x_{i,d}^H)^2} \quad (7)$$

979 Here,  $x_{i,d}^M$  and  $x_{i,d}^H$  denote the scores assigned by the model-based and human judge respectively.  
 980 RMSE quantifies the deviation between a model-based judge’s scores and those of human judge  
 981 in dimension  $d$ . Lower RMSE values indicate higher alignment with human preferences, and thus  
 982 higher reliability of the model-based judge.  
 983

984 **Human Agreement within Tolerance (A@ $\tau$ )** While RMSE measures the magnitude of error, it  
 985 can be sensitive to outliers. To provide a complementary view of judge reliability, we also introduce  
 986 the Human Agreement within Tolerance (A@ $\tau$ ) metric. This metric calculates the percentage of  
 987 evaluations where the judge’s score falls within a specified tolerance margin,  $\tau$ , of the human score.  
 988 Given the subjective nature of the scoring task, we set a tolerance of  $\tau = 1$  point. The metric, A@1,  
 989 is calculated as:  
 990

$$991 \quad 992 \quad 993 \quad 994 \quad \text{A@1} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(|x_{i,d}^M - x_{i,d}^H| \leq 1) \quad (8)$$

995 where  $\mathbb{I}(\cdot)$  is the indicator function, which is 1 if the condition is true and 0 otherwise. A higher  
 996 A@1 score indicates that the judge’s scores are more frequently in close agreement with human  
 997 evaluators, reflecting greater practical reliability.

998 This evaluation framework provides a comprehensive analysis of the performance of all the genera-  
 999 tors across multiple dimensions, ensuring objective comparison from both human and model-based  
 1000 perspectives.

### 1001 D.3 MODELS USED IN THE SEIR METHOD

1002 To construct the dataset, we adopt Qwen2.5-32B-Instruct as the language model ( $\mathcal{M}_L$ ), In-  
 1003 ternVL2.5-26B as the vision-language model ( $\mathcal{M}_V$ ), and FLUX.1-dev as the text-to-image gen-  
 1004 eration model ( $\mathcal{M}_G$ ). These open-source models are used as the default configuration in our SEIR  
 1005 framework. Importantly, the framework is modular by design—each component can be substituted  
 1006 with other models, offering flexibility for different deployment environments or research needs.

1007 To assess the generality and robustness of SEIR, we systematically evaluated its performance across  
 1008 a range of model configurations. Specifically, we experimented with different combinations of open-  
 1009 source LMs (InternLM, DeepSeek-R1, Qwen) and VLMs (InternVL, QwenVL), while keeping the  
 1010 generative model fixed as Flux. In addition, to benchmark against high-performing closed-source  
 1011 alternatives, we included a configuration that uses GPT-4o as both the LM and VLM, and DALL-E3  
 1012 as the generative model.

1013 The results, presented in Table 8, demonstrate several key trends:  
 1014

- 1015 • **SEIR consistently improves data quality.** Across all configurations, we observe notice-  
 1016 able improvements in all four evaluation dimensions—TCC, ICC, IQ and ITS—after ap-  
 1017 plying SEIR. The gains are particularly significant in the TCC, ICC and ITS dimensions,  
 1018 reflecting SEIR’s ability to enhance the semantic alignment and cooperative informative-  
 1019 ness of multimodal outputs.
- 1020 • **Initial quality varies across model combinations.** Among the open-source configura-  
 1021 tions, those involving Qwen (e.g., Qwen+InternVL or QwenVL) generally exhibit stronger  
 1022 performance in the No\_SEIR stage. In contrast, InternLM+QwenVL and DeepSeek-R1 +  
 1023 InternVL show relatively weaker initial consistency, suggesting differences in language-  
 1024 vision alignment quality across model families.

- **SEIR is especially effective for lower-performing combinations.** The relative improvements are more pronounced for model combinations with lower baseline performance. For example, InternLM + QwenVL improves by 0.60 in TCC and 0.70 in ITS, and DeepSeek-R1+InternVL shows notable gains across all metrics despite its initially modest performance.
- **SEIR narrows the gap between open- and closed-source models.** While the GPT-4o+DALL-E3 configuration achieves the highest initial quality across all metrics, the application of SEIR allows open-source configurations to reach comparable performance levels. For instance, Qwen+InternVL + Flux achieves 4.42 (TCC) and 4.51 (ITS) after SEIR, which closely rivals the 4.44 and 4.54 obtained by GPT-4o+DALL-E3.
- **Text-Image Synergy (ITS) shows the largest variance.** This dimension benefits the most from SEIR optimization, particularly in cases where image-text redundancy or disconnection was prevalent before refinement. The improvement indicates SEIR’s effectiveness in jointly adjusting both modalities to produce more complementary multimodal answers.

Overall, these results confirm that SEIR is a robust and generalizable enhancement framework. It consistently improves dataset quality across a wide range of model backbones, and significantly reduces the reliance on expensive closed-source models. Consequently, we adopt the open-source setup of Qwen+InternVL+Flux as our default configuration, balancing quality, flexibility, and cost-effectiveness. During the dataset generation and experimental testing process, this work consumed approximately 90,000 H100 hours.

Table 8: Comparison of dataset quality before and after SEIR optimization using different model configurations.

method	LM	VLM	GM	TCC	ICC	IQ	ITS
No_SEIR	InternLM	InternVL	Flux	3.72	3.98	4.30	3.70
SEIR	InternLM	InternVL	Flux	4.34	4.35	4.38	4.42
No_SEIR	InternLM	QwenVL	Flux	3.66	3.89	4.28	3.68
SEIR	InternLM	QwenVL	Flux	4.26	4.32	4.35	4.38
No_SEIR	DeepSeek-R1	InternVL	Flux	3.65	3.82	4.37	3.66
SEIR	DeepSeek-R1	InternVL	Flux	4.20	4.28	4.43	4.19
No_SEIR	Qwen	QwenVL	Flux	3.80	3.93	4.35	3.75
SEIR	Qwen	QwenVL	Flux	4.37	4.40	4.36	4.52
No_SEIR	Qwen	InternVL	Flux	3.85	4.01	4.42	3.79
SEIR	Qwen	InternVL	Flux	4.42	4.47	4.44	4.51
No_SEIR	GPT-4o	GPT-4o	DALL-E3	4.05	4.08	4.41	3.94
SEIR	GPT-4o	GPT-4o	DALL-E3	4.44	4.46	4.43	4.54

#### D.4 HYPERPARAMETERS USED FOR TRAINING JUDGE MODEL

We fine-tuned two large multimodal models, InternVL2.5-8B and QwenVL2.5-7B, as judge models to evaluate the interleaved image-text content completeness. We followed common practices for large scale model fine-tuning, applying weight decay regularization, learning rate warmup, and gradient clipping to ensure training stability. All experiments were conducted using mixed-precision training on distributed GPU clusters.

For the QwenVL2.5-7B judge model, we adopted a multi-GPU training setup using 4 devices with a total training batch size of 8, obtained by setting a per-device batch size of 1 and a gradient accumulation step of 2. The model was trained using the AdamW optimizer with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ , and  $\epsilon = 1 \times 10^{-8}$ . The initial learning rate was set to  $1 \times 10^{-5}$  and scheduled using a cosine decay strategy with a warmup ratio of 10%. A fixed random seed of 42 was used for reproducibility. Evaluation was conducted using a batch size of 8 per device, resulting in a total evaluation batch size of 32.

For the InternVL2.5-8B referee model, we adopted a multi-GPU training setup, using 4 devices with a total training batch size of 8. This was achieved by setting the batch size per device to 1 and the gradient accumulation steps to 2. The model was trained using the AdamW optimizer, with

$\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ ,  $\epsilon = 1 \times 10^{-8}$ , a gradient clipping threshold of 1.0, and a weight decay of 0.05. The initial learning rate was set to  $4 \times 10^{-5}$ , and a cosine annealing strategy was employed for adjustment, with a warm-up ratio of 3%. For reproducibility, a fixed random seed of 42 was used. During evaluation, the batch size per device was 8, resulting in a total evaluation batch size of 32.

### D.5 FURTHER VERIFICATION OF SEIR EFFECTIVENESS

To investigate the impact of the SEIR method, we conduct a fine-grained ablation by varying AR and IR iterations. Results in Table 9 show that both AR and IR contribute positively to performance. Specifically, increasing AR iterations mainly improves TCC, ICC, and ITS, while IR iterations further enhance ICC and ITS. These results confirm that AR enriches textual content and coherence, while IR reinforces visual relevance and multimodal synergy.

Table 9: Ablation study: impact of AR and IR iterations on Anole and VILA-U. “null” indicates baseline performance without SEIR-based training.

(a) Performance of Anole.						(b) Performance of VILA-U.					
AR	IR	TCC	ICC	IQ	ITS	AR	IR	TCC	ICC	IQ	ITS
null	null	3.09	3.01	2.92	2.26	null	null	2.46	3.72	3.37	2.19
0	0	3.33	3.17	2.92	2.77	0	0	3.07	3.83	3.37	3.13
1	1	3.37	3.21	3.07	2.71	1	0	3.1	3.8	3.34	3.2
2	2	3.41	3.27	3.03	2.79	2	0	3.17	3.79	3.39	3.27
3	0	3.47	3.20	3.03	2.85	3	0	3.17	3.81	3.36	3.27
3	1	3.51	3.25	3.11	2.93	3	1	3.17	3.81	3.4	3.3
3	2	3.49	3.3	3.08	2.91	3	2	3.21	3.85	3.37	3.29
3	3	3.52	3.32	3.1	2.94	3	3	3.19	3.83	3.39	3.33

### D.6 SYNJUDGE TRAIN/TEST DATA

Our full annotated data contains 48,000 samples. These 48,000 questions were generated by the SEIR method through three iterations, with the question template and topic hierarchy ensuring sufficient diversity across topics and conversational query style.

To obtain a broad distribution of multimodal question–answer (QA) outputs, these 48,000 questions were randomly assigned to different generators, which then produced responses. This strategy ensures that the resulting QA pairs cover a wide quality spectrum. After generation, all question–answer pairs were annotated by trained human annotators for TCC, ICC, IQ, and ITS over a two-week period.

The annotated dataset was split into a training set (80%) and a test set (20%). The training set was used to fine-tune candidate judge models (e.g., QwenVL\_trained, InternVL\_trained), while the test set was reserved for evaluation. For the evaluation, we used RMSE as the metric to measure how closely the scores from different judges align with ground-truth human annotations: a lower RMSE indicates a better judge. This evaluation process is what leads to the selection of our final model, SynJudge.

### D.7 DETAILED ANALYSIS OF THE JUDGE’S SCORING

**Judge Deviation Analysis.** To better understand each judge’s scoring behavior, we report the distribution of absolute score differences in Tables 10–14. Each table shows, for a given judge, the proportion of samples where the model’s score differs from the human reference by 0 to 5. These detailed distributions provide a fine-grained view of the judges’ alignment with human evaluators.

**Judge Agreement Analysis.** Table 15 provides a detailed breakdown of the **Human Agreement within Tolerance (A@1)** for each judge across all evaluated generators and dimensions. As formally defined in Appendix §D.2, this metric reflects the proportion of scores where the absolute difference between the model and human judge is no more than one point ( $\tau = 1$ ), which we consider an acceptable margin for subjective tasks.

1134 The results reinforce our findings from the main paper. The finetuned judges consistently achieve  
1135 higher agreement rates. QwenVL\_trained achieves the highest average A@1 of 95.4%. In-  
1136 ternVL\_trained also shows strong performance at 94.5%, significantly better than the non-finetuned  
1137 baselines. In contrast, zero-shot models like GPT-4o and QwenVL exhibit noticeably lower  
1138 agreement, around 86.5% and 87.5% respectively, indicating that they are less reliable for fine-  
1139 grained evaluation without specialized tuning. These findings further justify our selection of  
1140 QwenVL\_trained as the backbone for **SynJudge**.

1141

1142

1143

1144

1145

1146

1147

1148

1149

1150

1151

1152

1153

1154

1155

1156

1157

1158

1159

1160

1161

1162

1163

1164

1165

1166

1167

1168

1169

1170

1171

1172

1173

1174

1175

1176

1177

1178

1179

1180

1181

1182

1183

1184

1185

1186

1187

Table 10: Gap proportion between GPT-4o and human scoring.

Model	Dimension	Score					
		0	1	2	3	4	5
Anole	TCC	0.463	0.343	0.159	0.031	0.004	0.0
	ICC	0.525	0.325	0.106	0.037	0.007	0.0
	IQ	0.52	0.309	0.128	0.036	0.007	0.0
	ITS	0.536	0.289	0.111	0.056	0.007	0.001
GPT-4o+DALL-E	TCC	0.654	0.318	0.025	0.003	0.0	0.0
	ICC	0.686	0.276	0.031	0.003	0.002	0.002
	IQ	0.636	0.295	0.065	0.002	0.0	0.002
	ITS	0.635	0.303	0.039	0.019	0.002	0.002
DDiT	TCC	0.938	0.024	0.017	0.012	0.009	0.0
	ICC	0.467	0.379	0.111	0.037	0.006	0.0
	IQ	0.434	0.404	0.135	0.023	0.004	0.0
	ITS	0.843	0.043	0.023	0.022	0.027	0.042
Emu3	TCC	0.5	0.398	0.085	0.016	0.001	0.0
	ICC	0.535	0.333	0.082	0.037	0.012	0.001
	IQ	0.546	0.367	0.074	0.012	0.001	0.0
	ITS	0.454	0.358	0.123	0.052	0.013	0.0
SEIR	TCC	0.589	0.329	0.066	0.013	0.002	0.001
	ICC	0.686	0.251	0.037	0.017	0.009	0.0
	IQ	0.665	0.281	0.051	0.001	0.001	0.001
	ITS	0.577	0.3	0.063	0.037	0.016	0.007
Gemini+Flux	TCC	0.636	0.311	0.037	0.011	0.003	0.002
	ICC	0.705	0.237	0.036	0.015	0.007	0.0
	IQ	0.695	0.256	0.047	0.002	0.0	0.0
	ITS	0.657	0.267	0.045	0.023	0.006	0.002
Janus-Pro	TCC	0.31	0.419	0.234	0.036	0.001	0.0
	ICC	0.412	0.395	0.124	0.06	0.009	0.0
	IQ	0.428	0.369	0.164	0.034	0.005	0.0
	ITS	0.369	0.407	0.131	0.078	0.014	0.001
Liquid	TCC	0.465	0.389	0.12	0.024	0.002	0.0
	ICC	0.513	0.368	0.073	0.039	0.006	0.001
	IQ	0.478	0.399	0.094	0.025	0.002	0.002
	ITS	0.357	0.414	0.14	0.071	0.016	0.002
Show-o	TCC	0.525	0.343	0.104	0.026	0.002	0.0
	ICC	0.475	0.367	0.101	0.043	0.013	0.001
	IQ	0.445	0.382	0.139	0.029	0.005	0.0
	ITS	0.471	0.331	0.114	0.06	0.019	0.005
Show-o-Turbo	TCC	0.432	0.34	0.17	0.056	0.002	0.0
	ICC	0.53	0.332	0.095	0.032	0.011	0.0
	IQ	0.433	0.383	0.16	0.019	0.005	0.0
	ITS	0.457	0.36	0.122	0.044	0.013	0.004
VARGPT	TCC	0.351	0.356	0.235	0.051	0.007	0.0
	ICC	0.861	0.064	0.037	0.03	0.005	0.003
	IQ	0.875	0.082	0.034	0.003	0.003	0.003
	ITS	0.864	0.071	0.032	0.024	0.008	0.001
VILA-U	TCC	0.617	0.249	0.105	0.024	0.004	0.001
	ICC	0.428	0.405	0.117	0.046	0.004	0.0
	IQ	0.403	0.37	0.172	0.047	0.007	0.001
	ITS	0.562	0.254	0.1	0.038	0.026	0.02

Table 11: Gap proportion between InternVL and human scoring.

Model	Dimension	Score					
		0	1	2	3	4	5
Anole	TCC	0.523	0.294	0.147	0.034	0.001	0.001
	ICC	0.574	0.286	0.106	0.03	0.003	0.001
	IQ	0.504	0.272	0.176	0.042	0.005	0.001
	ITS	0.592	0.26	0.116	0.022	0.01	0.0
GPT-4o+DALL-E	TCC	0.658	0.326	0.015	0.001	0.0	0.0
	ICC	0.709	0.248	0.037	0.003	0.001	0.002
	IQ	0.647	0.269	0.075	0.006	0.0	0.003
	ITS	0.653	0.281	0.041	0.021	0.0	0.004
DDiT	TCC	0.942	0.022	0.02	0.005	0.01	0.001
	ICC	0.574	0.327	0.081	0.015	0.003	0.0
	IQ	0.485	0.347	0.136	0.029	0.003	0.0
	ITS	0.927	0.025	0.026	0.01	0.01	0.002
Emu3	TCC	0.586	0.329	0.075	0.008	0.002	0.0
	ICC	0.629	0.27	0.072	0.022	0.006	0.001
	IQ	0.531	0.328	0.12	0.021	0.0	0.0
	ITS	0.496	0.351	0.105	0.044	0.004	0.0
SEIR	TCC	0.589	0.352	0.048	0.01	0.001	0.0
	ICC	0.708	0.239	0.039	0.012	0.001	0.001
	IQ	0.692	0.246	0.057	0.003	0.001	0.001
	ITS	0.611	0.271	0.075	0.031	0.009	0.003
Gemini+Flux	TCC	0.629	0.332	0.032	0.004	0.002	0.001
	ICC	0.705	0.241	0.045	0.005	0.004	0.0
	IQ	0.694	0.248	0.056	0.002	0.0	0.0
	ITS	0.633	0.279	0.066	0.018	0.002	0.002
Janus-Pro	TCC	0.37	0.348	0.235	0.046	0.001	0.0
	ICC	0.52	0.327	0.124	0.026	0.003	0.0
	IQ	0.378	0.31	0.238	0.066	0.008	0.0
	ITS	0.471	0.313	0.161	0.049	0.006	0.0
Liquid	TCC	0.438	0.389	0.139	0.03	0.003	0.001
	ICC	0.552	0.339	0.076	0.03	0.003	0.0
	IQ	0.42	0.401	0.137	0.033	0.006	0.003
	ITS	0.433	0.343	0.149	0.059	0.013	0.003
Show-o	TCC	0.592	0.314	0.074	0.02	0.0	0.0
	ICC	0.556	0.319	0.1	0.02	0.004	0.001
	IQ	0.473	0.308	0.181	0.032	0.006	0.0
	ITS	0.49	0.319	0.129	0.05	0.008	0.004
Show-o-Turbo	TCC	0.528	0.333	0.118	0.02	0.001	0.0
	ICC	0.581	0.316	0.082	0.017	0.004	0.0
	IQ	0.447	0.324	0.199	0.023	0.006	0.001
	ITS	0.492	0.347	0.113	0.038	0.007	0.003
VARGPT	TCC	0.271	0.368	0.287	0.062	0.009	0.003
	ICC	0.857	0.069	0.057	0.011	0.003	0.003
	IQ	0.857	0.083	0.043	0.011	0.003	0.003
	ITS	0.872	0.051	0.055	0.014	0.007	0.001
VILA-U	TCC	0.675	0.238	0.069	0.015	0.002	0.001
	ICC	0.553	0.316	0.099	0.028	0.004	0.0
	IQ	0.408	0.303	0.205	0.076	0.005	0.003
	ITS	0.684	0.195	0.087	0.023	0.009	0.002

Table 12: Gap proportion between QwenVL and human scoring.

Model	Dimension	Score					
		0	1	2	3	4	5
Anole	TCC	0.783	0.189	0.027	0.001	0.0	0.0
	ICC	0.558	0.258	0.092	0.06	0.022	0.01
	IQ	0.577	0.294	0.102	0.022	0.005	0.0
	ITS	0.592	0.189	0.121	0.061	0.032	0.005
GPT-4o+DALL-E	TCC	0.679	0.299	0.022	0.0	0.0	0.0
	ICC	0.716	0.234	0.041	0.006	0.001	0.002
	IQ	0.579	0.325	0.081	0.012	0.003	0.0
	ITS	0.571	0.296	0.094	0.029	0.004	0.006
DDiT	TCC	0.947	0.021	0.013	0.008	0.01	0.001
	ICC	0.512	0.339	0.107	0.035	0.007	0.0
	IQ	0.484	0.317	0.17	0.029	0.0	0.0
	ITS	0.939	0.021	0.017	0.008	0.013	0.002
Emu3	TCC	0.628	0.316	0.05	0.006	0.0	0.0
	ICC	0.561	0.312	0.039	0.041	0.031	0.016
	IQ	0.58	0.345	0.073	0.002	0.0	0.0
	ITS	0.543	0.277	0.118	0.043	0.015	0.004
SEIR	TCC	0.574	0.339	0.076	0.009	0.002	0.0
	ICC	0.67	0.256	0.051	0.013	0.007	0.003
	IQ	0.667	0.264	0.063	0.003	0.001	0.002
	ITS	0.566	0.283	0.082	0.042	0.017	0.01
Gemini+Flux	TCC	0.607	0.354	0.031	0.004	0.003	0.001
	ICC	0.548	0.358	0.04	0.027	0.011	0.016
	IQ	0.677	0.254	0.064	0.003	0.002	0.0
	ITS	0.613	0.325	0.042	0.012	0.004	0.004
Janus-Pro	TCC	0.433	0.343	0.196	0.027	0.001	0.0
	ICC	0.454	0.29	0.166	0.054	0.025	0.011
	IQ	0.429	0.35	0.161	0.053	0.006	0.001
	ITS	0.502	0.276	0.132	0.059	0.028	0.003
Liquid	TCC	0.621	0.298	0.064	0.015	0.002	0.0
	ICC	0.571	0.311	0.089	0.024	0.004	0.001
	IQ	0.656	0.283	0.044	0.013	0.004	0.0
	ITS	0.478	0.239	0.115	0.067	0.09	0.011
Show-o	TCC	0.542	0.35	0.088	0.016	0.004	0.0
	ICC	0.536	0.316	0.105	0.036	0.004	0.003
	IQ	0.457	0.364	0.143	0.029	0.006	0.001
	ITS	0.571	0.26	0.094	0.051	0.021	0.003
Show-o-Turbo	TCC	0.436	0.336	0.187	0.038	0.003	0.0
	ICC	0.538	0.311	0.114	0.029	0.007	0.001
	IQ	0.554	0.343	0.078	0.024	0.001	0.0
	ITS	0.469	0.306	0.137	0.045	0.032	0.011
VARGPT	TCC	0.707	0.241	0.044	0.006	0.002	0.0
	ICC	0.858	0.069	0.046	0.015	0.009	0.003
	IQ	0.852	0.073	0.052	0.015	0.005	0.003
	ITS	0.832	0.054	0.057	0.036	0.018	0.003
VILA-U	TCC	0.823	0.151	0.017	0.005	0.003	0.001
	ICC	0.443	0.377	0.149	0.031	0.0	0.0
	IQ	0.447	0.356	0.15	0.041	0.006	0.0
	ITS	0.678	0.165	0.104	0.034	0.013	0.006

Table 13: Gap proportion between InternVL\_trained and human scoring.

Model	Dimension	Score					
		0	1	2	3	4	5
Anole	TCC	0.871	0.106	0.023	0.0	0.0	0.0
	ICC	0.738	0.193	0.05	0.012	0.007	0.0
	IQ	0.703	0.206	0.068	0.019	0.003	0.001
	ITS	0.773	0.161	0.052	0.013	0.001	0.0
GPT-4o+DALL-E	TCC	0.801	0.191	0.008	0.0	0.0	0.0
	ICC	0.767	0.211	0.019	0.001	0.002	0.0
	IQ	0.777	0.205	0.014	0.001	0.002	0.001
	ITS	0.778	0.201	0.013	0.004	0.003	0.001
DDiT	TCC	0.967	0.013	0.012	0.007	0.001	0.0
	ICC	0.77	0.186	0.034	0.006	0.004	0.0
	IQ	0.72	0.215	0.045	0.019	0.001	0.0
	ITS	0.962	0.015	0.014	0.007	0.001	0.001
Emu3	TCC	0.775	0.2	0.021	0.004	0.0	0.0
	ICC	0.747	0.196	0.037	0.016	0.004	0.0
	IQ	0.748	0.2	0.038	0.009	0.001	0.004
	ITS	0.741	0.185	0.039	0.031	0.004	0.0
SEIR	TCC	0.726	0.25	0.017	0.006	0.001	0.0
	ICC	0.816	0.159	0.022	0.002	0.0	0.001
	IQ	0.737	0.241	0.008	0.002	0.002	0.01
	ITS	0.732	0.233	0.028	0.005	0.001	0.001
Gemini+Flux	TCC	0.786	0.195	0.012	0.004	0.0	0.003
	ICC	0.728	0.243	0.02	0.007	0.002	0.0
	IQ	0.746	0.236	0.016	0.002	0.0	0.0
	ITS	0.806	0.142	0.028	0.02	0.001	0.003
Janus-Pro	TCC	0.855	0.133	0.012	0.0	0.0	0.0
	ICC	0.725	0.209	0.057	0.008	0.001	0.0
	IQ	0.699	0.226	0.052	0.019	0.004	0.0
	ITS	0.679	0.219	0.081	0.018	0.001	0.002
Liquid	TCC	0.846	0.138	0.012	0.004	0.0	0.0
	ICC	0.796	0.169	0.021	0.012	0.001	0.001
	IQ	0.716	0.224	0.052	0.006	0.001	0.001
	ITS	0.71	0.205	0.066	0.017	0.002	0.0
Show-o	TCC	0.788	0.176	0.023	0.01	0.003	0.0
	ICC	0.743	0.205	0.028	0.019	0.004	0.001
	IQ	0.718	0.215	0.059	0.006	0.001	0.001
	ITS	0.676	0.214	0.082	0.021	0.004	0.003
Show-o-Turbo	TCC	0.718	0.213	0.054	0.013	0.002	0.0
	ICC	0.684	0.243	0.052	0.015	0.006	0.0
	IQ	0.713	0.242	0.024	0.017	0.004	0.0
	ITS	0.715	0.214	0.06	0.01	0.001	0.0
VARGPT	TCC	0.797	0.15	0.047	0.006	0.0	0.0
	ICC	0.913	0.04	0.032	0.008	0.004	0.003
	IQ	0.922	0.033	0.03	0.011	0.001	0.003
	ITS	0.917	0.027	0.038	0.016	0.001	0.001
VILA-U	TCC	0.906	0.078	0.012	0.002	0.002	0.0
	ICC	0.719	0.208	0.062	0.01	0.001	0.0
	IQ	0.702	0.216	0.059	0.022	0.001	0.0
	ITS	0.827	0.122	0.044	0.005	0.001	0.001

Table 14: Gap proportion between QwenVL\_trained and human scoring.

Model	Dimension	Score					
		0	1	2	3	4	5
Anole	TCC	0.857	0.121	0.022	0.0	0.0	0.0
	ICC	0.713	0.225	0.052	0.006	0.004	0.0
	IQ	0.701	0.229	0.05	0.016	0.004	0.0
	ITS	0.748	0.189	0.048	0.013	0.001	0.001
GPT-4o+DALL-E	TCC	0.783	0.207	0.01	0.0	0.0	0.0
	ICC	0.765	0.209	0.023	0.001	0.001	0.001
	IQ	0.757	0.227	0.015	0.0	0.001	0.0
	ITS	0.773	0.187	0.029	0.009	0.002	0.0
DDiT	TCC	0.974	0.015	0.007	0.003	0.001	0.0
	ICC	0.732	0.223	0.039	0.005	0.001	0.0
	IQ	0.733	0.227	0.033	0.004	0.003	0.0
	ITS	0.969	0.013	0.013	0.004	0.001	0.0
Emu3	TCC	0.757	0.22	0.021	0.002	0.0	0.0
	ICC	0.729	0.233	0.019	0.019	0.0	0.0
	IQ	0.749	0.224	0.021	0.005	0.001	0.0
	ITS	0.781	0.17	0.04	0.007	0.001	0.001
SEIR	TCC	0.699	0.279	0.018	0.003	0.001	0.0
	ICC	0.659	0.309	0.022	0.01	0.0	0.0
	IQ	0.704	0.281	0.014	0.001	0.0	0.0
	ITS	0.707	0.268	0.02	0.004	0.001	0.0
Gemini+Flux	TCC	0.735	0.236	0.022	0.004	0.003	0.0
	ICC	0.708	0.262	0.023	0.004	0.003	0.0
	IQ	0.712	0.263	0.024	0.001	0.0	0.0
	ITS	0.703	0.271	0.015	0.01	0.001	0.0
Janus-Pro	TCC	0.841	0.147	0.011	0.001	0.0	0.0
	ICC	0.672	0.262	0.054	0.012	0.0	0.0
	IQ	0.667	0.252	0.072	0.007	0.001	0.001
	ITS	0.7	0.224	0.064	0.01	0.001	0.001
Liquid	TCC	0.833	0.152	0.011	0.004	0.0	0.0
	ICC	0.704	0.245	0.03	0.018	0.002	0.001
	IQ	0.699	0.244	0.044	0.01	0.003	0.0
	ITS	0.661	0.269	0.06	0.009	0.001	0.0
Show-o	TCC	0.774	0.195	0.021	0.01	0.0	0.0
	ICC	0.688	0.247	0.043	0.022	0.0	0.0
	IQ	0.722	0.248	0.023	0.004	0.002	0.001
	ITS	0.693	0.244	0.051	0.01	0.001	0.001
Show-o-Turbo	TCC	0.69	0.246	0.047	0.016	0.001	0.0
	ICC	0.672	0.265	0.039	0.023	0.001	0.0
	IQ	0.624	0.299	0.068	0.008	0.001	0.0
	ITS	0.747	0.197	0.045	0.009	0.001	0.001
VARGPT	TCC	0.82	0.159	0.017	0.004	0.0	0.0
	ICC	0.913	0.047	0.031	0.009	0.0	0.0
	IQ	0.922	0.05	0.027	0.001	0.0	0.0
	ITS	0.925	0.038	0.019	0.016	0.001	0.001
VILA-U	TCC	0.89	0.097	0.012	0.001	0.0	0.0
	ICC	0.639	0.277	0.058	0.025	0.001	0.0
	IQ	0.646	0.272	0.072	0.01	0.0	0.0
	ITS	0.8	0.127	0.06	0.011	0.002	0.0

Table 15: Evaluation accuracy A@1 comparison across judges

Model	Dim.	GPT-4o	InternVL	InternVL_trained	QwenVL	QwenVL_trained
Anole	TCC	0.805	0.816	0.977	0.972	0.977
	ICC	0.848	0.859	0.930	0.816	0.937
	IQ	0.828	0.776	0.908	0.870	0.929
	ITS	0.823	0.851	0.933	0.779	0.936
GPT-4o+DALL-E	TCC	0.971	0.983	0.991	0.978	0.990
	ICC	0.961	0.956	0.977	0.949	0.974
	IQ	0.931	0.916	0.981	0.903	0.983
	ITS	0.938	0.933	0.978	0.866	0.959
DDiT	TCC	0.960	0.963	0.979	0.967	0.989
	ICC	0.846	0.901	0.955	0.850	0.954
	IQ	0.837	0.831	0.934	0.801	0.958
	ITS	0.885	0.951	0.976	0.958	0.982
Emu3	TCC	0.896	0.915	0.974	0.944	0.975
	ICC	0.867	0.899	0.942	0.871	0.961
	IQ	0.912	0.858	0.946	0.924	0.972
	ITS	0.811	0.846	0.925	0.819	0.950
SEIR	TCC	0.917	0.940	0.975	0.912	0.978
	ICC	0.936	0.946	0.973	0.925	0.967
	IQ	0.945	0.937	0.976	0.930	0.983
	ITS	0.876	0.881	0.964	0.848	0.974
Gemini+Flux	TCC	0.946	0.960	0.980	0.961	0.970
	ICC	0.941	0.945	0.970	0.905	0.969
	IQ	0.950	0.941	0.981	0.930	0.975
	ITS	0.924	0.911	0.946	0.936	0.973
Janus-Pro	TCC	0.727	0.716	0.987	0.775	0.987
	ICC	0.806	0.846	0.932	0.743	0.933
	IQ	0.796	0.687	0.924	0.778	0.918
	ITS	0.775	0.783	0.897	0.777	0.924
Liquid	TCC	0.853	0.827	0.983	0.918	0.984
	ICC	0.880	0.890	0.964	0.881	0.948
	IQ	0.877	0.820	0.939	0.939	0.942
	ITS	0.770	0.775	0.914	0.716	0.929
Show-o	TCC	0.867	0.905	0.963	0.891	0.969
	ICC	0.842	0.874	0.947	0.852	0.935
	IQ	0.827	0.781	0.932	0.820	0.969
	ITS	0.801	0.808	0.889	0.829	0.936
Show-o-Turbo	TCC	0.771	0.861	0.929	0.772	0.935
	ICC	0.861	0.895	0.926	0.848	0.936
	IQ	0.815	0.770	0.954	0.897	0.922
	ITS	0.816	0.838	0.928	0.775	0.943
VARGPT	TCC	0.706	0.638	0.946	0.948	0.978
	ICC	0.924	0.924	0.953	0.926	0.959
	IQ	0.956	0.939	0.953	0.924	0.971
	ITS	0.934	0.922	0.944	0.884	0.962
VILA-U	TCC	0.865	0.912	0.982	0.973	0.986
	ICC	0.832	0.868	0.926	0.818	0.915
	IQ	0.772	0.711	0.916	0.802	0.917
	ITS	0.815	0.879	0.948	0.842	0.926
Average	A@1	0.865	0.866	0.945	0.875	0.954

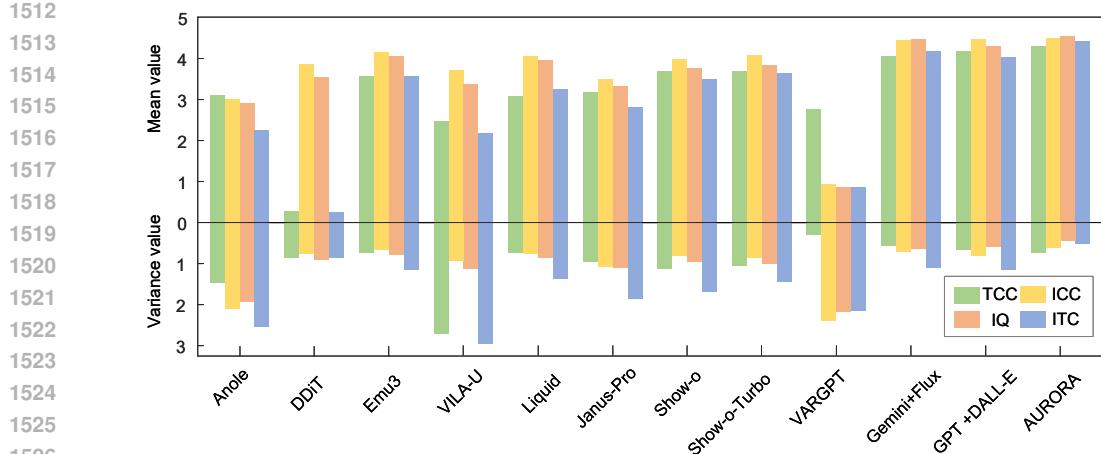


Figure 7: A visualization of mean and variance of different generators

#### 1530 D.8 DETAILED ANALYSIS OF THE CAPABILITIES OF DIFFERENT GENERATORS

1531  
**Mean and Variance Analysis.** Figure 7 presents the mean and variance of evaluation scores across  
1532 TCC, ICC, IQ, and ITC for different generators. Several important observations emerge from the  
1533 results. First, DDT exhibits the lowest mean scores in both TCC (0.38) and ITC (0.37), indicating  
1534 poor content coverage and weak image-text synergy. Second, VARGPT shows mean scores  
1535 below 1 across ICC, IQ, and ITC, suggesting significant deficiencies in visual generation capabilities  
1536 and multimodal alignment. Third, VILA-U demonstrates the highest variance in TCC and  
1537 ITC among all models, implying that its performance is highly unstable across different questions.  
1538 In contrast, Gemini-Flux and GPT-4o+DALL-E achieve mean scores above 4.0 across all evalua-  
1539 tion dimensions, reflecting generally strong performance. However, their relatively high variance  
1540 in ITC reveals that they still struggle with maintaining image-text consistency and complementarity  
1541 across samples. Most notably, SEIR method consistently outperforms all other generators across  
1542 all four dimensions, achieving the highest mean scores while maintaining the lowest variance. This  
1543 indicates not only superior quality but also high stability and robustness in both textual and visual  
1544 generation. These findings collectively highlight the importance of both quality and consistency for  
1545 robust multimodal generation, and demonstrate the synergy of SEIR in constructing high quality,  
1546 stable datasets.

#### 1548 E QUESTION TEMPLATES AND TOPIC HIERARCHY

1549  
1550 In our framework, **question templates** are designed to model the style of human queries rather  
1551 than to encode domain knowledge. Specifically, these templates capture the recurrent **syntactic and**  
1552 **pragmatic structures** through which users naturally formulate requests. For example, variations  
1553 such as “*Please recommend some equipment needed for hiking.*”, “*Tell me some equipment needed*  
1554 *for hiking.*”, or “*Could you tell me what equipment is needed for hiking?*” all convey the same under-  
1555 lying intent but differ in their linguistic form. By systematically enumerating such structures, we  
1556 obtain broad coverage of syntactic patterns for conversational queries (e.g., “can you ...”, “please  
1557 ...”), while leaving the semantic content to be drawn from a large topic hierarchy. This separation  
1558 ensures that the **linguistic diversity** of user queries can be effectively modeled: the question tem-  
1559 plate specifies *how* a request is asked, whereas the the 3500-topic hierarchy defines *what* the request  
1560 is about. The combination of these two components enables our dataset to achieve high diversity  
1561 while faithfully reflecting real-world user interaction styles.

#### 1562 E.1 QUESTION TEMPLATES

1563  
1564 Do you know \*\*\*? Can you draw a picture of it for me?  
1565 Do you know what \*\*\* looks like? Please draw an image of it for me.

1566 I'm very interested in \*\*\*. Please help me describe it and draw a  
 1567 portrait of it.  
 1568 Can you imagine what \*\*\* looks like? Please draw a portrait of it for me.  
 1569 Have you paid attention to \*\*\*? Can you tell me something about it?  
 1570 Besides, can you depict it in a painting?  
 1571 What do you think \*\*\* looks like? Can you draw what this place might look  
     like?  
 1572 What does \*\*\* look like? Can you draw a sketch of it for me?  
 1573 Do you know about \*\*\*? Please draw an image of them for me.  
 1574 I heard that \*\*\* is very attractive. Can you introduce it to me? Then  
     draw an image of it for me.  
 1575 I need a painting of \*\*\* now. Please help me describe it and draw it.  
 1576 I'm paying a lot of attention to \*\*\* now. Do you know it? By the way,  
     help me draw a picture to introduce it.  
 1577 Can you draw a picture of \*\*\*? Besides, can you give me some science  
     knowledge about \*\*\*?  
 1578 Have you ever seen the scene of \*\*\*? What does \*\*\* usually look like?  
 1579 Please draw a picture for me.  
 1580 Will there be \*\* in the \*\* of \*\*? Can you draw a picture of this scene  
     for me?  
 1581 What is the most wonderful \*\*\* you have ever seen? Please draw a scene  
     picture of \*\*\* for me.  
 1582 What kind of wonders can be seen in \*\*\*? I'm really curious about the  
     scene. Can you show it to me?  
 1583 Can you imagine \*\*\*, with \*\* and \*\*, creating a moment of \*\* and \*\*? Help  
     me draw them.  
 1584 Can you describe \*\*\* for me? It would be even better if there is a  
     painting.  
 1585 Please introduce \*\*\* and draw about it.  
 1586 Hey, can you tell me \*\*\* and draw a picture?  
 1587 ...  
 1588  
 1589  
 1590  
 1591  
 1592  
 1593  
 1594

## E.2 TOPIC HIERARCHY

1595 To provide a clearer picture of the dataset's composition, the table 16 below shows the distribution  
 1596 across our 8 primary domains, including the number of categories and fine-grained topics  
 1597 within each. As you can see, the data is well-distributed across diverse topics like Natural Scenery  
 1598 (19.88%), Cultural Scenery (15.87%), and Animals (14.77%), ensuring broad and deep coverage.  
 1599

1600 Table 16: Data distribution of InterSyn across 8 primary domains. The table details the percentage  
 1601 domain distribution (Domain Dist.), distribution per Category (Dist. per Cat.), number of categories  
 1602 (# Categories), and number of topics for each domain (# Topics), demonstrating broad and deep  
 1603 coverage.

Domain	Domain Dist. (%)	Dist. per Cat. (%)	# Categories	# Topics
Animals	14.77	1.64	9	517
Plants	10.13	1.27	8	355
Natural Scenery	19.88	2.48	8	695
Cultural Scenery	15.87	1.98	8	555
Objects	10.65	1.33	8	373
Activities	11.20	1.40	8	392
Food	6.71	0.84	8	235
Culture	10.80	1.35	8	378

1615 {  
 1616 "Animals": {  
 1617 "Terrestrial Animals": ["Giant panda", "Snow leopard", "Black bear", "Red  
     panda", "Tibetan antelope", "Argali" ...],  
 1618 "Marine Animals": ["Blue whale", "Killer whale", "Great white shark", "  
     Humpback whale", "Dolphin", "Octopus" ...],  
 1619

```

1620 "Extinct Animals": ["Dinosaur", "Dodo", "Woolly mammoth", "Saber-toothed
1621 tiger", "Pterosaur", "Stegosaurus"...,]
1622 "Domesticated Animals": ["Pet cat", "Pet dog", "Pet bird", "Mouse", "
1623 Ornamental fish", "Cow", "Sheep", "Pig"...,]
1624 },
1625 "Plants": {
1626 "Edible Plants": ["Rice", "Wheat", "Corn", "Sorghum", "Oat", "Buckwheat",
1627 "Quinoa", "Millet", "Barley"...,]
1628 "Medicinal Plants": ["Ginseng", "Wolfberry", "Coptis chinensis", "
1629 Notoginseng", "Astragalus membranaceus", "Angelica sinensis"...,]
1630 "Ornamental Plants": ["Rose", "Tulip", "Chrysanthemum", "Peony", "Chinese
1631 peony", "Lily", "Narcissus", "Hyacinth", "Iris"...,]
1632 },
1633 "Natural Scenery": {
1634 "Mountains & Forests": ["The flag cloud of Mount Everest in the Himalayas
1635 ", "Alpine meadows and wildflowers in the Alps"...,]
1636 "Water & Weather": ["Hawaiian volcanic lava flowing into the sea", "The
1637 blue-domed church in Santorini, Greece", "A dugout canoe in the lagoon of
1638 Tahiti"...,]
1639 "Deserts & Volcanoes": ["The sharp ridges on the backlit side of sand
1640 dunes", "The winding silhouette of a camel caravan"...,]
1641 "Seasons & Landforms": ["Red-crowned cranes dancing in the winter snow in
1642 Hokkaido, Japan", "The tulip maze in Keukenhof Gardens in the
1643 Netherlands in spring"...,]
1644 },
1645 "Cultural Scenery": {
1646 "Cities & Villages": ["The mirror-like water surface of the terraced
1647 fields in Yuanyang, Yunnan", "The lavender fields in Provence, France
1648 "...,]
1649 "Religion & Religious Sites": ["Devout believers praying in front of the
1650 Western Wall in Jerusalem", "The play of light and shadow under the dome
1651 of St..."],
1652 "Heritages & Wonders": ["The giant paintings of the Nazca Lines in Peru
1653 seen from above", "The Treasury at the end of the Siq in Petra, Jordan",
1654 "The Pyramids of the Sun and Moon in Teotihuacan,..."]
1655 },
1656 "Objects": {
1657 "Household & Daily Items": ["Bench", "Chair", "Sofa", "Coffee table", "
1658 Bookshelf", "Wardrobe", "Desk", "Dressing table", "Bed", "Dining table",
1659 "Dining chair"...,]
1660 "Military & Security": ["Pistol", "Rifle", "Submachine gun", "Machine gun
1661 ", "Artillery", "Missile", "Tank", "Armored vehicle", "Fighter jet"...,]
1662 "Tools & Equipment": ["Fire hydrant", "Wrench", "Screw", "Hammer", "
1663 Shovel", "Screwdriver", "Tape measure", "Electric drill", "Pliers", "Saw
1664 ", "File", "Soldering iron"...,]
1665 "Energy & Industry": ["Generator", "Solar panel", "Wind turbine", "
1666 Hydraulic generator", "Battery", "Inverter", "Transformer", "Charging
1667 pile", "Oil drum", "Gas cylinder"...,]
1668 "Transportation & Communication": ["Bicycle", "Car", "Motorcycle", "
1669 Airplane", "Bus", "Train", "Truck", "Ship", "Traffic light", "Tricycle",
1670 "Electric scooter"...,]
1671 },
1672 "Activities": {
1673 "Daily Life & Occupations": ["Doctor", "Firefighter", "Farmer", "Teacher
1674 ", "Lawyer", "Craftsman", "Researcher", "Photographer", "Singer", "Dancer
1675 ", "Painter", "Journalist"...,]
1676 "Emotional & Social Interactions": ["Hug", "Kiss", "Meet", "Talk", "
1677 Lecture", "Study", "Shake hands", "Comfort", "Celebrate", "Take a group
1678 photo", "Quarrel", "Share", "Wave"...,]
1679 "Sports & Labor": ["Run", "Play basketball", "Play football", "Play
1680 volleyball", "Play badminton", "Play tennis", "Play table tennis", "Jump
1681 ", "Ride a bike", "Box", "Wrestle"...,]
1682 },
1683 "Food": {

```

```

1674 "Regional Cuisines": ["Mapo Tofu", "Ramen", "Braised Pork Belly in Soy
1675 Sauce", "Scrambled Eggs with Tomatoes", "Shredded Pork with Green Peppers
1676 ", "Braised Beef with Potatoes"..."],
1677 "Baked Goods & Desserts": ["Caramel Pudding", "Macaron", "Donut", "Cake",
1678 "Yogurt", "French Croissant", "Italian Tiramisu", "German Black Forest
1679 Cake", "Japanese Wagashi"..."],
1680 "Processed Foods": ["Snacks", "Canned Food", "Frozen Food", "Biscuits", "Chocolate
1681 Biscuits", "Ice Cream", "Popcorn", "Potato Chips", "Canned Fish
1682 ", "Frozen Dumplings"..."],
1683 "Beverages": ["Red Wine", "Chinese Baijiu", "Beer", "Coke", "Juice", "Tea
1684 ", "Milk", "Soda Water", "French Champagne", "Italian Espresso", "Japanese
1685 Sake", "Korean Makgeolli"..."],
1686 "Pet Food": ["Dog Food", "Cat Food", "Chew Sticks", "Bones", "Pet Canned
1687 Food", "Freeze-Dried Chicken Pieces", "Salmon-Flavored Cat Treats"..."]
1688 },
1689 "Culture": {
1690 "Material Culture": ["Hanfu (Han Chinese Clothing)", "Qipao (Cheongsam)",
1691 "Kimono", "Indian Sari", "Western Suit", "Wedding Dress", "Tangzhuang (Tang-style
1692 Costume)", "Mongolian Robe"..."],
1693 "Spiritual Culture": ["The Dragon Totem in Ancient China", "The Phoenix
1694 Totem", "The Eagle Totem of the Native Americans", "The Wolf Totem", "The
1695 Rainbow Serpent Totem of the Australian Aborigines"..."],
1696 "Behavioral Culture": ["Traditional Chinese Wedding", "Western Church
1697 Wedding", "Coming-of-Age Ceremony", "Crowning Ceremony", "Sacrificial
1698 Ceremony", "Japanese Tea Ceremony Etiquette", "The Hongi (Nose Rubbing)
1699 of the Maori People", "The Namaste of India", "The Apprenticeship
1700 Ceremony in Thailand", "The Torch Festival Ceremony of the Yi Ethnic
1701 Group"..."]
1702 }
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727

```

## F EVALUATION DIMENSIONS FOR DATASET QUALITY

### F.1 QUESTION EVALUATION DIMENSIONS

**Reasonableness of Expression:** The question statement is smooth, without any grammatical errors, and the words are used accurately and appropriately. For example, "Please introduce the Great Wall to me and also give me a picture of the Great Wall" is a reasonable expression; while "Tell me about the Great Wall, and give me a picture" has a problem of confused expression. Such questions will affect the model's understanding of the intention and make it difficult to give an accurate answer.

**Clarity of Requirements:** Clearly indicate that the model is required to provide both text and image responses simultaneously. For instance, "Introduce the appearance characteristics of Notre-Dame de Paris and provide a high-definition frontal picture", which clearly puts forward the dual requirements of text description and image acquisition; if the question is just "What does Notre-Dame de Paris look like", without clearly stating the image requirement, it does not meet the requirements and cannot effectively guide the model to give a comprehensive response.

**Focus of the Theme:** The question revolves around a single and clear theme and will not jump between multiple unrelated themes. For example, "Introduce the geographical features of Mount Fuji and attach a distant view of Mount Fuji", with the theme focused on Mount Fuji; while "Tell me about Mount Fuji and then talk about the Eiffel Tower, and give two corresponding pictures", which involves two different themes, may lead to unclear logical answers from the model and is not conducive to the standardized construction of the dataset.

**Feasibility and Clarity:** Based on common sense judgment, the content involved in the question is something that the model has the ability to answer through language and images, and there is no way of multiple interpretations, and the model can accurately grasp the questioner's intention. For example, "Describe the living habits of giant pandas and give a picture of a panda eating bamboo", the model can answer based on its existing knowledge reserve and image generation ability, and the intention is clear; however, "Tell me what it's like for a person to take a bath in volcanic magma and give a picture", such questions seriously deviate from reality and lack scientific basis. The model can

1728 neither answer based on existing knowledge, nor is there a real-world reference for image generation,  
 1729 which will lead to absurd and meaningless generation results and greatly reduce the reliability and  
 1730 practicality of the dataset.

1731 **Appropriateness of Length:** The length of the question is moderate, which not only contains  
 1732 enough key information to guide the model to generate high quality answers but also is not too  
 1733 long and complicated for the model to grasp the key points. Generally speaking, short and concise  
 1734 questions are helpful for the model to quickly understand the intention, such as "Introduce the For-  
 1735 bidden City and give a panoramic picture"; but being too short may lack sufficient information, such  
 1736 as "Forbidden City, picture"; and overly long and cumbersome questions, like "Please introduce  
 1737 in detail the process of the changes of the Forbidden City since its construction in the Ming Dy-  
 1738 nasty through various dynasties, including the evolution of architectural styles, the transformation  
 1739 of functional uses and other aspects, and provide a high-definition panoramic picture that can com-  
 1740 prehensively display the current overall layout of the Forbidden City. At the same time, ensure that  
 1741 the picture contains the main palaces, courtyards, city walls and other iconic elements of the Forbid-  
 1742 den City", may cause confusion when the model processes it. The ideal length can be determined  
 1743 according to practical experience and testing. Usually, about 15 - 50 words is more appropriate,  
 1744 which can convey the requirements completely and also facilitate the model to process efficiently.

## 1745 F.2 INTERLEAVED IMAGE-TEXT ANSWER EVALUATION DIMENSIONS

1746 **Text Content Completeness (0-5 points):** This dimension only focuses on the correspondence  
 1747 between the text response and the question, whether the content precisely matches the user's needs,  
 1748 and whether the information is complete and error-free. It does not consider the output of any other  
 1749 dimensions and evaluation criteria.

1750 0 points: No text appears; 1 point: The text answer has nothing to do with the question; it is  
 1751 completely wrong, completely divorced from the question, and there is no positive response to the  
 1752 text requirement; there is less content but there are truncations and random spitting characters. 2  
 1753 points: The text answer can only cover a small part of the elements required in the question, and there  
 1754 is a large amount of unreasonable content; there is a very obvious phenomenon of text truncation  
 1755 that seriously affects the original information; the content is very long or very short, which seriously  
 1756 affects the reading. 3 points: The answer can correspond to key elements, there is a small amount of  
 1757 unreasonable content, and there may be omissions of key information; the content is too long or too  
 1758 short, but the information basically corresponds. 4 points: The required elements of the question are  
 1759 basically all corresponding, there is no unreasonable content, there is a omission of key information,  
 1760 or the answer is awkward; the content is slightly longer or shorter, but the answer is very correct.  
 1761 5 points: The content of the answer exactly corresponds to the question, there is no unreasonable  
 1762 content, and the answer is smooth and fluent, with full content.

1763 **Image Content Completeness (0-5 points):** This dimension only focuses on the correspondence  
 1764 between the image content and the question (considering the content of the picture, the degree to  
 1765 which the image content answers the question). Whether the key parts are retained, and whether  
 1766 there is an obvious lack of objects.

1767 0 points: No image appears; 1 point: The content of the image is completely wrong, and no key  
 1768 elements are depicted at all; the image has no connection to the problem, even if the image itself  
 1769 is of good quality. 2 points: About half of the key elements required for the problem are missing,  
 1770 and there are a large number of unreasonable elements; the elements in the figure may have some  
 1771 connection to the problem, but it is almost impossible to identify what they are. 3 points: Only  
 1772 a small number of key elements required for the problem are missing in the figure, most of the  
 1773 elements can be fully identified, and there are only a few unreasonable content. 4 points: Basically  
 1774 lack the elements required for the problem, and there may be minor flaws in some details. 5 points:  
 1775 All the elements required for the question are completely corresponding, the main body is intact,  
 1776 and the picture content answers the question very well.

1777 **Image Quality (0-5 points):** This dimension only focuses on the performance of the basic genera-  
 1778 tion technology of the image (do not consider the content of the picture). Whether it is clear, whether  
 1779 there are blurred, noisy or out-of-focus areas, truncations or damages (that is, the judgment of image  
 1780 aesthetics and subjective quality).

1781 0 points: No picture; 1 point: The image is very ugly, and it is almost impossible to identify the  
 1782 image content. 2 points: The image looks ugly, the overall image is blurred but can be barely recog-

1782 nized; 3 points: The image is medium in appearance, and the main elements can be distinguished,  
 1783 but other elements are blurred. 4 points: The image looks good, the picture is relatively clear, and  
 1784 there is no visible blurring phenomenon; 5 points: The image looks good, the details are sharp  
 1785 without blur, and the image quality is very high.

1786 **Image-Text Synergy (0–5 points):** This dimension evaluates the degree of alignment and comple-  
 1787 mentarity between the textual and visual components of a response. It focuses not only on how well  
 1788 the entities or scenes described in the text are accurately and completely depicted in the image, but  
 1789 also on whether the text and image together form a coherent and mutually supportive answer to the  
 1790 question.

1791 0 points: The image and text are completely unrelated. Additionally, if either the image or the text  
 1792 is missing (i.e., “null”), the response is assigned 0 points. 1 point: The image and text are minimally  
 1793 related, with only a few elements weakly corresponding. The response lacks coherence and fails to  
 1794 effectively address the question. 2 points: Around half of the key elements described in the text are  
 1795 reflected in the image, but significant mismatches remain. The overall synergy is poor. 3 points:  
 1796 Most elements between the text and image are consistent, but a few important mismatches or omis-  
 1797 sions in key entities or scenes reduce the completeness of the response. 4 points: Nearly all elements  
 1798 between the text and image are consistent, with only minor mismatches in non-critical details. The  
 1799 response answers the question well, but there may be redundancy between the two modalities, lim-  
 1800 iting their complementarity. 5 points: The text and image are perfectly aligned, with all described  
 1801 elements accurately and fully presented. The two modalities work together in a complementary way  
 1802 to form a complete and informative response without unnecessary duplication.

## 1804 G ALL PROMPTS USED IN THIS WORK

### 1806 G.1 PROMPTS USED IN SEIR METHOD

1808 Only a simple example of a single-round dialogue generation prompt is provided here. The most  
 1809 detailed prompts are given in detail in the open-source code. Detailed prompts can be found in our  
 1810 code.

1811 Here is the prompt for the question generation:

```

1813 I am building a question-answer dataset.
1814 The topic of this dataset item is {{topic}}. Your task is to generate a
1815 question based on this topic.
1816 The length of the question should not exceed 50 words. Here is the
1817 question template: \n{ques_temp}.\n
1818 The new question you generate can refer to the sentence pattern of the
1819 question template.
1820 The question must meet the following detailed requirements:
1. **Incorporate Image Request Naturally**: Clearly express the need to
1821 generate a picture, but use varied and creative expressions to make the
1822 request feel natural and human-like. Avoid repetitive phrases like 'maybe
1823 generate a picture.' Instead, use diverse sentence structures to request
1824 the image.
2. **Varied Sentence Structures**: Diversify how questions are phrased.
1825 Use different ways of asking, such as open-ended questions, hypothetical
1826 scenarios, or requests for examples.
3. **Conciseness and Clarity**: Ensure the question is still concise and
1827 immediately understandable but without sounding repetitive or formulaic.
1828 Avoid redundant language.
4. **Topic Relevance**: Keep the question focused on the given topic {{topic}},
1829 ensuring it remains engaging and meaningful. Avoid weak
1830 connections to the topic.
5. **Approachable Tone**: Use a conversational, approachable tone that
1831 mimics real human interactions. Keep it friendly and engaging, avoiding
1832 overly formal or robotic expressions.
6. **Lexical Simplicity with Creativity**: Use everyday vocabulary with
1833 occasional creative language that fits the topic. Ensure accessibility
1834 for a broad audience while maintaining interest.
  
```

1836 7. **\*\*Question Value and Inspiration\*\*:** Make the question thought-  
 1837 provoking or creative, capable of inspiring meaningful answers. Avoid  
 1838 overly simple or overly complex questions.  
 1839 8. **\*\*Image Context\*\*:** Clearly specify what kind of picture is expected,  
 1840 but do so creatively.  
 1841 Output only the generated question directly. Do not include explanations,  
 1842 instructions, or any extra text.

1843 Here is the prompt to get the the suggestions for the question:  
 1844

1845 I am currently constructing a question-answer dataset. The first step is  
 1846 to imitate human needs and tone based on a certain topic and ask a  
 1847 question.  
 1848 This question needs to include the requirement for generating textual  
 1849 content and a picture.  
 1850 The topic is: ({topic}) .\n  
 1851 The following is a question generated based on this topic: \n{old\_q}\n  
 1852 You need to analyze the quality of this question from a human perspective  
 1853 , such as whether the question is too wordy?  
 1854 Is the question sentence pattern not commonly used in human daily  
 1855 communication? How well does the question fit the topic?  
 1856 Does the tone of the question sound human? Are there any uncommon  
 1857 expressions in the sentence?  
 1858 Is it a meaningless question? Does the question contain a request for  
 1859 generating an image? Is the generated question easy to answer? And so on.  
 1860 You need to help me provide revision suggestions. It would be best if the  
 1861 suggestions are concise and brief, and not too long.  
 1862 If you think the original question is not good in other aspects, you need  
 1863 to help me give modification suggestions.  
 1864 Only output the modification suggestions in the end, and there is no need  
 1865 to output the modified results.  
 1866 Your output should conform to this format {json\_format}  
 1867 If you think the original question is good enough, you don't need to give  
 1868 improvement suggestions. You only need to output None.  
 1869 Therefore, your final output is either None or the modification  
 1870 suggestions.

1871 Here is the prompt for the question modification:  
 1872

1873 I am currently constructing a question-answer dataset.\n  
 1874 The following is the original question generated by an LLM: \n{old\_q}\n\n  
 1875 However, I believe the quality of this question can be improved, as it  
 1876 doesn't sound like something people would naturally ask in daily  
 1877 communication.\n\n  
 1878 I have provided some modification suggestions: {mod\_q\_suggestion}.\n  
 1879 Please revise the question based on these suggestions and the given topic  
 1880 , making it sound more natural and human-like.\n  
 1881 Finally, output only the modified question without any additional text.

1882 Here is the prompt to get the answer of the question:  
 1883

1884 Currently, I'm constructing a question-answer dataset. This is the  
 1885 current question: \n{final\_q}\n  
 1886 Since this question usually contains a requirement for textual answer and  
 1887 image generation., but you don't need to generate the actual image.  
 1888 Instead, you should generate an answer and a description of the image  
 1889 according to the question.  
 1890 To ensure high Image-Text Synergy (ITS), write the answer line so it  
 1891 gives the core explanation while referencing key visual elements, and  
 1892 write the caption line so it adds complementary details that the text  
 1893 omits; the two lines must stay tightly aligned, avoid duplication, and  
 1894 together convey more than either could alone.  
 1895 Therefore, your response should include an answer to the question: answer  
 1896 ; and a description of the image: caption. And you are not allowed to

1890 output responses like 'I can't generate images.' You need to pretend that  
 1891 you can.  
 1892 The image description must not exceed 65 words. This last point is very  
 1893 important! You just need to output in two lines and there should be no  
 1894 other content. The output content: start the first line with 'answer:',  
 1895 representing the answer; start the second line with 'caption:',  
 1896 representing the caption.  
 1897 Your answer should be related to the previous content and must not be  
 1898 repetitive.

1899 Here is the prompt for the suggestions for answer modification:

1900 Currently, I'm constructing a question-answer dataset.  
 1901 Here is the question: \n{final\_q}\n.  
 1902 The question usually includes the requirement for textual answer and  
 1903 image generation.  
 1904 Then, here is the answer to this question:\n{old\_ac}\n  
 1905 The answer is divided into two parts, including the textual answer to the  
 1906 question and an image description.  
 1907 Do you think the combination of this answer and image description can  
 1908 fully meet the requirements of the question? Are the image description  
 1909 and the answer content consistent and not redundant?  
 1910 How is the correlation among the question, the answer and the image  
 1911 description? Does it conform to the habits of human answering questions?  
 1912 If you were a knowledgeable human expert, how do you think you would  
 1913 answer this question? Would the answer seem too wordy?  
 1914 Would the overlap between the answer and the image description be too  
 1915 high? Can the image description well summarize a picture?  
 1916 If you were a nitpicking critic, do you think there are areas for  
 1917 improvement in this question, the answer and the image description?  
 1918 Would the image description be too short and not rich enough in content?  
 1919 Are there any discriminatory elements in the answer and the image  
 1920 description? And so on.  
 1921 You can give modification suggestions based on the above aspects. Or if  
 1922 you think the answer is unreasonable in other aspects, you also need to  
 1923 give your modification suggestions.  
 1924 In addition, the modification suggestions need to be divided into two  
 1925 parts: the answer and the image description. And the content needs to be  
 1926 concise and condensed, not overly long.  
 1927 Or if you think the answer and the image description are already perfect,  
 1928 you don't need to put forward improvement suggestions, and just output  
 1929 None.  
 1930 Therefore, your final output is either None or the modification  
 1931 suggestions.  
 1932 Only output the modification suggestions in the end, and there is no need  
 1933 to output the modified results.Your output should conform to this format  
 1934 {json\_format} or None.

1935 Here is the prompt for the answer modification:

1936 You are tasked with improving the output of a model output based on the  
 1937 suggestion feedback.  
 1938 Here is the context and what you need to do step by step:\n\n  
 1939 Model Output to Modify (old\_ac): \n{old\_ac}\n  
 1940 This is the current answer generated by the model. The answer is divided  
 1941 into two parts:\n  
 1942 - 'answer': This is the text answer to the question.\n  
 1943 - 'caption': This is the image description associated with the answer.\n\n  
 1944 Modification Suggestion (mod\_ac\_suggestion): \n{mod\_ac\_suggestion}\n  
 1945 This is the suggestion for improving the model's output, including  
 1946 corrections or enhancements to both the 'answer' and 'caption' parts.\n\n  
 1947 Your task is to:\n  
 1948 - According to the provided mod\_ac\_suggestion, update the 'answer' and '  
 1949 caption' sections in old\_ac.\n

1944 - Ensure that the updated 'caption' does not exceed 65 words.\n  
 1945 - Follow the specified format strictly.\n\n  
 1946 Important: You just need to output in two lines and there should be no  
 1947 other content.  
 1948 The output content: start the first line with 'answer:', representing the  
 1949 answer; start the second line with 'caption:', representing the caption.

1950 Here is the prompt for the suggestions for caption modification:

1952 Currently, I'm constructing a question-answer dataset.  
 1953 Here is the question: \n{final\_q}\n.  
 1954 This question usually contains a request for generating textual content  
 and a picture.  
 1955 Then, this is the original answer final\_a: {final\_a} and the image  
 1956 description old\_c: {old\_c} generated according to this question,  
 1957 You now need to evaluate the quality of the image description and the  
 1958 image based on the question and the answer. Does the image match the  
 1959 image description?  
 1960 When proposing revisions, follow the Image-Text Synergy (ITS) principle:  
 1961 suggest changes that make the picture (and its caption) complement rather  
 1962 than repeat the fixed textual answer, depict the visual elements the  
 1963 answer references, reduce redundancy or irrelevant details, and keep full  
 1964 factual consistency so that image+text together convey more than either  
 1965 could alone.  
 1966 How is the degree of correlation between the image description and the  
 1967 content of the answer to the question? Can the image description well  
 1968 summarize the content of the picture?  
 1969 Are there any unreasonable objects or behaviors in the image? Is the  
 1970 image description clear and not wordy? And so on.  
 1971 You can give modification suggestions regarding the image description  
 1972 based on the above aspects. Suggestions in other aspects not mentioned  
 1973 above are also highly encouraged to be put forward.  
 1974 The revision suggestions you provide need to be concise and condensed,  
 1975 and shouldn't be too long.  
 1976 If you think the image description and the image for this question and  
 1977 answer are already perfect, then you don't need to put forward any  
 1978 suggestions and just output None.  
 1979 Therefore, your final output is only None or the modification suggestions  
 .  
 1980 Only output the modification suggestions in the end, and there is no need  
 1981 to output the modified results.Your output should conform to this format  
 1982 {json\_format} or None.

1983 Here is the prompt for the caption modification:

1984 Currently, I'm constructing a question-answer dataset. The question  
 1985 usually includes the requirement for textual answer and image generation.  
 1986 Then, this is the image description of the answer: \n{old\_c}\n  
 1987 Then I think the quality of the image description is not very high.  
 1988 I have provided some modification suggestions here: \n{mod\_c\_suggestion}\n  
 1989 Please regenerate the image description according to these suggestions.  
 1990 The length of the picture description should not exceed 65 words. In the  
 1991 end, you only need to output the modified image description.

## 1992 G.2 INTERLEAVED IMAGE-TEXT ANSWER EVALUATION PROMPT USED BY MLLM

1993 Here is the prompt for evaluating the interleaved image-text answer:

1994 You are an experienced, fair and impartial judge. Next, I will provide  
 1995 you with a conversation where a human interacts with different GPTs on  
 1996 daily topics. In this scenario, the human will pose a text question, and  
 1997 the GPT's response is based on this question. This response usually  
 1998 includes a piece of text and image information, but there may be

1998 exceptions where there is only text or only image information. Now you  
 1999 need to reasonably rate the response given by the GPT. <chatbegin>  
 2000 represents the start of the Q&A data, and <chatend> represents the end of  
 2001 the Q&A data. The rating of the response is divided into the following  
 2002 four dimensions, and you should rate the response fairly and impartially  
 2003 according to the criteria of each dimension.

2004 Here are the four dimensions for evaluating the response:  
 2005 " "  
 2006 <Interleaved Image-Text Answer Evaluation Dimensions>  
 2007 " "  
 2008 The content of your output rating must strictly conform to the following  
 2009 format:  
 2010 [Text Content Completeness: \*; Image Content Completeness: \*; Image  
 2011 Quality: \*; Image-Text Synergy: \*]  
 2012 your score \* for different dimensions, only as a score in (0, 1, 2, 3, 4,  
 2013 5). You need to strictly conduct the grading.  
 2014 Here is the data you need to evaluate, and you need to evaluate the  
 2015 quality of the Answer from the above four dimensions (both text and image  
 2016 may be "null", and the fact that one of them is "null" will not affect  
 2017 the rating of other dimensions.):

## H BENCHMARK SAMPLES

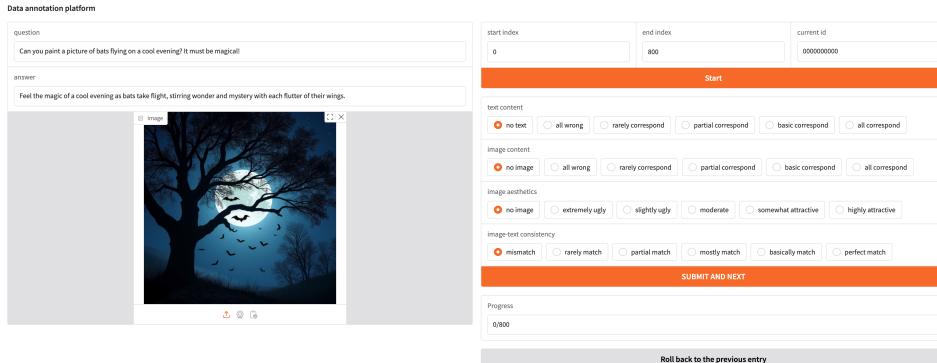
2018 The partly benchmark examples obtained after modification based on the questions raised by the  
 2019 participants are as follows:

2020 Are there fireflies in the forest on a summer night? Can you draw a  
 2021 picture of this scene for me?  
 2022 What could a wonderful concert scene be like? Please draw a scene of a  
 2023 concert for me.  
 2024 What kind of wonders can be seen in the forest on a cold winter night? I'm  
 2025 very curious about what that scene would be like. Can you show it to me  
 2026 ?  
 2027 Who is the king of the African savanna? Can you draw a picture to depict  
 2028 it?  
 2029 Can you imagine a serene ocean scene with a setting sun and some seagulls  
 2030 , creating a calm and relaxing moment? Please draw it for me.  
 2031 Can you describe a forest for me? It would be even better if there is a  
 2032 painting.  
 2033 Hey, can you tell me a really terrifying legend and draw a vivid picture  
 2034 of it?  
 2035 Can you describe the scenes in a futuristic music video? If possible, can  
 2036 you quickly draw a sketch? I'm really eager to see your ideas!  
 2037 Hey! What is the daily life of people in the military usually like? Also,  
 2038 can you show me what a soldier in military uniform looks like?  
 2039 Describe the scene of a huge lightning bolt during a storm. Draw a  
 2040 picture of this scene.  
 2041 Can you imagine how the concept of the Tree of Life is presented in  
 2042 different religions? Perhaps a painting showing its symbolism would be  
 2043 helpful.  
 2044 Can you describe what the snowy scene in a blizzard is like? I want to  
 2045 see such a landscape.  
 2046 Introduce a delicious snack. Describe its appearance, ingredients, and  
 2047 what makes it so appealing. Also, draw what it looks like.  
 2048 Describe the traditional decorations of the Lantern Festival and show me  
 2049 a picture of a lively lantern display.  
 2050 Can you describe an autumn scene with vivid orange-red leaves under a  
 2051 clear blue sky? Then draw an image of it.  
 2052 What are some interesting behaviors of cats? Can you show me a picture of  
 2053 a cat marking its territory?  
 2054 Can you quickly draw a picture of the Christ the Redeemer statue in Rio  
 2055 Janeiro and share some interesting facts about it?

2052 Hey! Can you describe a spring garden scene? I really want to hear enough  
 2053 details, and you need to draw it according to the description!  
 2054 Do you know what happened 100 years ago? Please draw a history-related  
 2055 picture! Thank you!  
 2056 I need a picture of the age of the dinosaurs now. Do you know about past  
 2057 history? Please draw a picture for me.  
 2058 Can you describe what a basketball court is like? Draw a basketball  
 2059 moment for me.  
 2060 What is the Lantern Festival like? Can you show me some pictures of  
 2061 traditional lanterns?  
 2062 What kind of casual outfit do you think is suitable for wearing on a  
 2063 relaxed Saturday afternoon? Can you draw what it looks like?  
 2064 Can you describe and perhaps draw a picture showing a person practicing  
 2065 yoga in a tranquil park at sunrise?  
 2066 I need a landscape picture of the countryside. Please describe it and  
 2067 draw an image for me.  
 2068 Can you draw a picture of an airplane for me? Also, give me some popular  
 2069 science knowledge about it.  
 2070 ...

## I HUMAN ANNOTATION PLATFORM

2072 We develop a human annotation platform to evaluate the quality of interleaved image-text responses.  
 2073 Annotators assess each response across four predefined dimensions, focusing on the content and  
 2074 coherence between visual and textual elements. To ensure annotation reliability, cross-validation is  
 2075 conducted on high-rated samples. An overview of the annotation interface is shown in Figure 8.



2089 Figure 8: Human annotation platform to evaluate the quality of interleaved image-text responses.  
 2090

## J LIMITATIONS OF THIS STUDY

2094 While our work introduces InterSyn, the first large scale, instruction-following dataset for multi-  
 2095 turn, interleaved image-text dialogues, and proposes SynJudge, a comprehensive automatic eval-  
 2096 uator emphasizing image-text synergy, several limitations remain that suggest directions for future  
 2097 improvement.

2098 First, although our SEIR framework substantially enhances output quality through multi-stage re-  
 2099 finement, the visual fidelity of generated images is inherently constrained by the upper bounds of  
 2100 current text-to-image models. This may limit the expressiveness and precision of visual responses,  
 2101 particularly for fine-grained or specialized topics.

2103 Second, our current dataset is restricted to one image per dialogue turn, which simplifies the mod-  
 2104 eling process but diverges from real-world scenarios where understanding or generating multiple  
 2105 images simultaneously is often necessary—e.g., comparative reasoning, procedural steps, or spa-  
 2106 tial reasoning tasks. While we have experimentally validated the feasibility of multi-image dialogue

2106 generation using alternative synthesis pipelines, such functionality is not yet reflected in the released  
2107 dataset.

2108 Third, the SynJudge evaluator is currently designed to assess single-image responses, meaning it  
2109 does not fully capture the additional complexity and multimodal dependencies introduced by multi-  
2110 image contexts. Extending SynJudge to support multi-image evaluation is a promising future direc-  
2111 tion.

2112 Finally, although InterSyn spans diverse domains and fine-grained topics, future work could enhance  
2113 its coverage of highly structured tasks or multi-modal reasoning chains that involve deeper world  
2114 knowledge or long-term dialogue coherence.

2115 These limitations highlight important opportunities for scaling interleaved image-text datasets and  
2116 improving evaluators toward more generalizable, high-fidelity multimodal generation systems.

2118

2119

2120

2121

2122

2123

2124

2125

2126

2127

2128

2129

2130

2131

2132

2133

2134

2135

2136

2137

2138

2139

2140

2141

2142

2143

2144

2145

2146

2147

2148

2149

2150

2151

2152

2153

2154

2155

2156

2157

2158

2159