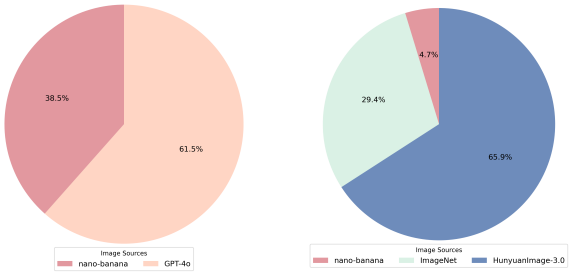
UniVerse: Empower Unified Generation with Reasoning and Knowledge

Supplementary Material

A. Dataset Statistics

The distribution of image sources for the Reasoning and Knowledge subsets of our dataset is illustrated in Figure 5. The Reasoning subset comprises images generated by two T2I models: Nano-Banana [11] and GPT-4o [22]. For Nano-Banana, images are synthesized using explicit prompts derived from pre-generated text triples. The GPT-4o images are sourced from the existing Echo-4o dataset [40], where each image was processed by an LLM to firstly produce a description, then assign a sub-category, and finally generate corresponding text triples based on the templates we provide.

The Knowledge subset includes images from a combination of synthetic and real-world sources. Synthetic images are generated using the HunyuanImage-3.0 [2] and Nano-Banana T2I models, while real-world images are obtained from ImageNet, all of which pertain to the Entity Knowledge category. This multi-source approach, particularly the inclusion of real-world images, mitigates the risk of model overfitting to the stylistic biases of any single T2I model.



(a) Reasoning Subset (b) Knowledge Subset

Figure 5. Distribution of Image Sources

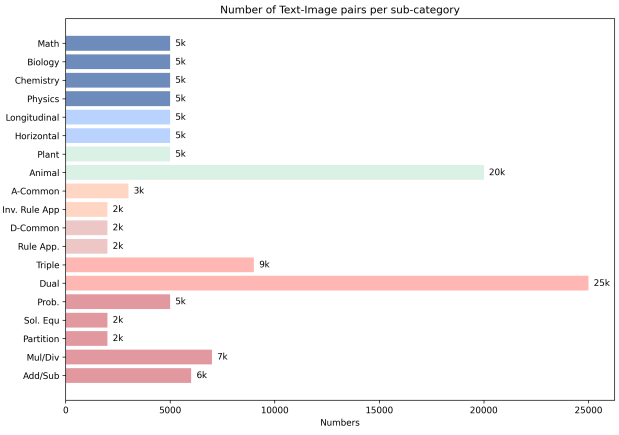


Figure 6. Number of Text-Image pairs per sub-category

The number of text-image pairs for each subcategory is shown in Figure 6. Most of the images from GPT-4o are categorized to Dual-Binding subcategory of Spatial-Attribute Constraint Reasoning, so this category is the largest portion, with 25k samples.

B. Data Filtering

Here we explain what retain and what we discard in the data filtering stage. *What we retain*: SOTA T2I models (Nano-banana) may fail to render prompts with precise counts (e.g. 6 potted plants) or complex relationships (bowls of different colors in a row RRYBBB), producing images with incorrect object numbers (7 instead of 6) or arrangements (RYRBBB). Instead of discarding these hard examples, we simply adjust the number or relationship in the implicit prompt and reasoning chain to make them match the image. Examples are shown in Figure 7. This kind of 'post-hoc' reasoning maintains the original causal link without disrupting it. *What we discard*: The T2I models are unsta-

Category	Original Implicit Prompt	Adjusted Implicit Prompt
Arithmetic Add/Sub	A windowsill has 2 potted plants. Someone adds 4 more. Now draw the potted plants on the windowsill.	A windowsill has 3 potted plants. Someone adds 4 more. Now draw the potted plants on the windowsill.
Spatial-Attr Dual-binding	6 bowls of red, yellow and blue colors are arranged in a row from left to right. 1 is yellow, the number of blue bowls is 1 more than the red. All blue are at the right end, all red are at the left end...	6 bowls of red, yellow and blue colors are arranged in a row from left to right. 1 is yellow, the number of blue bowls is 1 more than the red. All blue are at the right end, the yellow one is between 2 red bowls...

Figure 7. Illustration of how the implicit prompt is adjusted when chosen for retention during the data filtering stage.

ble when generating images that depict scientific domains, e.g., chemical reactions, often requiring many attempts to achieve high quality. We used a template to generate a large number of prompts and images for these categories and simply discarded the inaccurate ones. As shown in Table 3 below, the Arithmetic category yielded the most retained images. The Discipline category had the highest proportion of discarded images.

Table 3. Proportion of retained and discarded images for different categories.

Category	Arithmetic	Spatial-Attr	Deduct.	Abduct.	Discipline	Spatio-Temp	Entity(ImageNet)
One-Round Pass	67.8%	70.5%	85.2%	79.8%	65.2%	92.6%	100%
Salvaged	**24.4%**	21.3%	9.2%	17.1%	0%	7.3%	0%
Discard	7.9%	8.3%	5.6%	3.1%	**34.8%**	0.2%	0%

C. More Experiment Results

C.1. Experiment Result on different models

To demonstrate the efficacy of our dataset across different architectures, we train another two UMMs, Show-o [37]

and BLIP3o-8B [3], using the *without CoT* training methodology. Table 4 presents the evaluation results on the WISE benchmark, comparing the baseline performance of these models with their performance when trained without CoT-Injection. Both models show improvements after training with our data, which proves that our dataset can generalize across different UMMs and enhance their reasoning capabilities.

Table 4. **Benchmark Evaluation Results.** This table includes the evaluation results of the two baseline models (Show-o and BLIP3o-8B), and these models trained without CoT-Injection (*w/o CoT*) on WISE benchmarks. Higher value indicates better.

Model	WISE						
	CL	TM	SP	BIO	PH	CH	**Overall**
Unified Multimodal Models							
Show-o	0.28	0.36	0.40	0.23	0.33	0.22	0.30
Show-o$_{w/o\ CoT}$	0.34	0.39	0.46	0.23	0.36	0.24	0.34
BLIP3o-8B	0.49	0.51	0.63	0.54	0.63	0.37	0.52
BLIP3o-8B$_{w/o\ CoT}$	0.54	0.55	0.67	0.55	0.64	0.37	0.56

C.2. Warm-up Ablation

The warm-up ablation study is shown in the Table 5 below. We evaluate the model performances with different warm-up phases on the WISE benchmark, and empirically find that 20% yields the best score.

Warm-up phase	0%	10%	20%	30%	40%
WISEScore	0.69	0.72	0.74	0.71	0.71

Table 5. Warm-up Ablation Study

D. Human Validation

To prevent bias and errors, we define fine-grained topics in subcategories, manually write seed prompts, and employ diverse LLMs and T2I models (as detailed in Section 3.2). We conducted a human validation on a sample of 3000 images. On a 5-point scale, annotators rated the factual correctness of the text triplets (implicit-reasoning-explicit) and image alignment. The high scores, shown in Table 6, confirm our dataset's reliability.

Human Evaluation	Text factual correctness	Image alignment
Score	4.32	4.19

Table 6. Human evaluation scores