FROM KNOWLEDGE GAPS TO BETTER GENERATION: AUDITING AND COMPLETING WORLD KNOWLEDGE IN TEXT-TO-IMAGE TRAINING DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image (T2I) models achieve high visual quality but remain inconsistent in depicting concepts grounded in world knowledge, particularly those sparsely represented in training data. Because datasets are typically organized by source, concept coverage gaps are difficult to identify and address. We present CGTR, a framework that maps large scale image–text data to a hierarchical world knowledge graph for coverage analysis and targeted training-data refinement. Auditing 370M annotation records from publicly available T2I data reveals missing and underrepresented concepts across semantic levels. These statistics guide the construction of KnowGapBench, which covers 3,286 underrepresented concepts and evaluates concept identity separately from compositional compliance. CGTR independently audits each target training dataset to identify supplementation targets. World knowledge and diverse scene graph templates guide caption construction, while curated reference images guide appearance during image synthesis. Caption consistency and image–text alignment checks filter the resulting training pairs. The hierarchy organizes original and supplementary data into semantic groups, and concept-aware sampling regulates their exposure during training. We apply CGTR to Fine-T2I and fine-tune UniWorld-V1 with the refined data under its original generative objective. The overall Knowledge score on KnowGapBench increases from 15.3 to 22.5, while performance on familiar concepts outside the supplementation target set is preserved.

1 INTRODUCTION

Recent text-to-image (T2I) models have made substantial progress in visual quality, prompt adherence, and stylistic control, supporting applications in graphic design, advertising, and visual storytelling (Wu et al., 2025; Z-Image Team, 2025). However, public image–text corpora have highly imbalanced concept distributions, and models often generate infrequent concepts less reliably than frequent ones (Parashar et al., 2024). Increasing data volume or average quality does not ensure adequate training support for every concept. Reliable concept generation also requires diverse examples that consistently associate each concept with its defining visual features (Huang et al., 2026; 2025b).

A central difficulty is the mismatch between source-level data organization and concept-level learning. When training data are organized and mixed by source, evidence for the same concept can remain fragmented across names, aliases, and annotation schemes. Source-level weights do not directly specify concept-level exposure, making it difficult to inspect the data supporting a failed concept or adjust the mixture for a chosen concept family. In large-scale visual learning, hierarchical clustering mitigates imbalance through balanced sampling in feature space (Vo et al., 2024), but its unlabeled clusters do not themselves provide canonical concept identities or explicit controls over named targets. Recent graph-based data engines offer more direct semantic control (Z-Image Team, 2025). We connect such semantic organization to cross-source coverage measurements that inform data supplementation, training mixtures, and evaluation.

We present CGTR (Coverage-Guided Training Refinement), a framework that connects concept coverage auditing with training-data refinement and evaluation. We curate concepts and factual relations from Wikipedia and Wikidata, remove redundant entries, and organize visually relevant

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

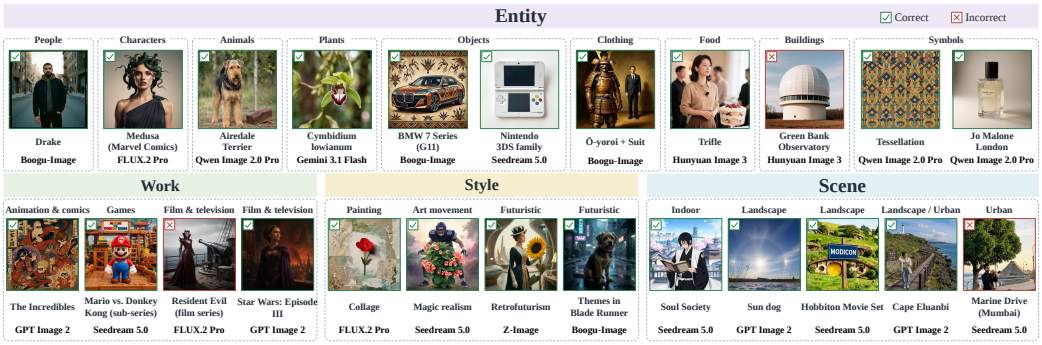


Figure 1: KnowGapBench examples spanning entities, works, styles, and scenes. Target concepts and generating models are indicated. Green checks and red crosses denote passes and failures in concept identity.

concepts into a hierarchical world knowledge graph. This graph provides a shared concept space for linking heterogeneous annotations to canonical identities across datasets, names, and aliases. We audit 370M annotated image–text records from 20 public datasets and additional web and synthetic collections (Fang et al., 2026; Dai et al., 2025; Shen et al., 2026). Despite the corpus scale, the most frequent 1% of linked concepts account for 94.98% of total concept support. These measurements expose coverage gaps and concentration that source-level dataset statistics cannot resolve.

These findings motivate evaluating T2I models on concepts with limited support in the audited corpus. Existing world-knowledge benchmarks select concepts based on visual specificity or semantic domains (Huang et al., 2026; 2025b; Niu et al., 2026), but these criteria do not directly reflect their representation in training data. We construct KnowGapBench by selecting concepts from the curated concept space that are unobserved or sparsely supported in the audited corpus. This provides a quantitative, traceable basis for target selection, complementing existing curation criteria (Table 1). Figure 1 presents representative examples from KnowGapBench, covering entities, works, styles, and scenes.

To improve concept generation, CGTR combines targeted data supplementation with concept-aware sampling. Reweighting cannot provide missing visual evidence, while unconstrained synthesis may introduce implausible scenes or alter compositional constraints (Gao et al., 2026). Limited supervision can also entangle concept identity with recurring contexts (Yang et al., 2026). Therefore we audit each target training dataset separately and supplement its missing and underrepresented concepts. One-hop factual context from the curated knowledge graph and diverse scene-graph templates guide caption construction, while curated reference images guide concept appearance during synthesis. Caption consistency and image–text alignment checks filter supplementary pairs, which are combined with original data resampled using category priors. Applied to Fine-T2I (Ma et al., 2026), CGTR raises UniWorld-V1's (Lin et al., 2025) overall Knowledge score on KnowGapBench from 15.3 to 22.5 while preserving performance on common concepts outside the supplementation targets.

Our key contributions are summarized as follows:

- We construct a hierarchy of 254,067 visually relevant canonical concepts and audit 370M annotated image–text records through a shared mapping procedure, making missing support and long-tailed imbalance measurable across data sources and semantic levels.
- We develop CGTR to synthesize diverse examples for concepts with insufficient training support using factual context and curated references, and regulate their exposure through concept-aware sampling.
- We introduce KnowGapBench, containing 3,286 concepts in 30 categories and 3,698 prompts across four tasks, from concept realization to joint concept composition. The benchmark selects targets directly from measured coverage gaps and evaluates concept correctness separately from compositional compliance.

Table 1: Comparison of visual-concept benchmarks. Cat.: categories; Joint: multi-target concept tasks; Ref.: paired target images; Freq.: quantitative frequency analysis. Category counts follow each benchmark's taxonomy; dashes denote unreported counts or protocols.

Benchmark	Concept inventory		Evaluation coverage			Construction and scoring		
	Cat.	Concepts	Joint	Ref.	Freq.	Source	Selection basis	Scores
EntityDrawBench	5	250	—	✓	✓	Wiki. / GLD / Fandom	Category sampling	Faithfulness / realism
KITTEN	8	322	—	✓	—	OVEN-Wiki	Entity-domain curation	Entity / instruction
T2I-FactualBench	8	1,600	✓	✓	—	CNER / BabelNet	Lexical specificity	Concept / composition
Rare-Concept-1K	5	1,000	—	—	—	Wiki. / domain texts	Expert verification	Attribute / context
TextTIGER	2	—	—	✓	—	PopVQA / Wikipedia	Source availability	Entity / text
NICE	5	—	✓	—	—	Web topics	Novel/trending entities	Prompt / preference
KnowGapBench	**30**	**3,286**	✓	✓	✓	Wikipedia	Measured coverage	Knowledge / Atomic

2 Related Work

Text-to-image model evaluation. Compositional benchmarks test objects, counting, attributes, and relations (Ghosh et al., 2023; Huang et al., 2025a). Knowledge benchmarks examine entity fidelity (Chen et al., 2022; Huang et al., 2026), rare, Wikipedia-linked, or emerging concepts (Venkatesh et al., 2024; Ozaki et al., 2025; Son et al., 2025), and implicit world knowledge (Niu et al., 2026; Zhang et al., 2025a;b). T2I-FactualBench (Huang et al., 2025b) uses multi-round VQA to evaluate factuality and composition across concept memorization, contextual instantiation, and multi-concept composition. Fine-grained evaluation uses question dependencies, VQA alignment scores, and element-level human annotations (Cho et al., 2024; Lin et al., 2024; Han et al., 2026). KnowGap-Bench selects targets from measured corpus coverage gaps (Table 1), separating reference-assisted identity judgments from compositional scores.

Data synthesis and training strategies. Filtering and semantic balancing support large-scale image–text curation (Gadre et al., 2023; Fang et al., 2024; Xu et al., 2024; Vo et al., 2024), while synonym-aware frequency analysis links sparse training support to rare-concept failures (Parashar et al., 2024). Qwen-Image combines filtering, balancing, and targeted acquisition (Wu et al., 2025); Z-Image (Z-Image Team, 2025) uses a hierarchical knowledge graph to identify underrepresented concepts and rebalance training data. Recaptioning and iterative synthetic-data selection improve supervision (Li et al., 2025c; Sun et al., 2025). Generate Any Scene (Gao et al., 2026) samples objects, attributes, and relations into scene graphs of controlled complexity, then derives captions and evaluation questions. Other synthesis pipelines target uncommon scenarios, generation diversity, subject consistency, and scientific accuracy (Ye et al., 2025; Fang et al., 2026; Kumari et al., 2025; Li et al., 2025b). CGTR uses dataset-specific coverage gaps to guide scene-graph synthesis, with factual context and visual references preserving concept identity. The same hierarchy supports corpus analysis and concept-aware sampling.

3 Hierarchical World Knowledge Graph Construction and Concept Coverage Analysis

We organize visually relevant concepts and factual relations from Wikipedia and Wikidata into a hierarchical graph. We map the aggregated image–text corpus, $\mathcal{D}_{\text{corpus}}$, to measure coverage for benchmark construction, and each training dataset separately to guide retrieval, supplementation, and concept-aware sampling.

3.1 Hierarchical World Knowledge Graph Construction

Concept collection and filtering. We collect concepts with their English and Chinese names and aliases, merge duplicates, and remove disambiguation and administrative pages (Figure 2(a)). We retain factual predicates relevant to visual generation and discard metadata and maintenance relations. After retaining the top 60% of concepts by PageRank (Brin & Page, 1998), we assess visual

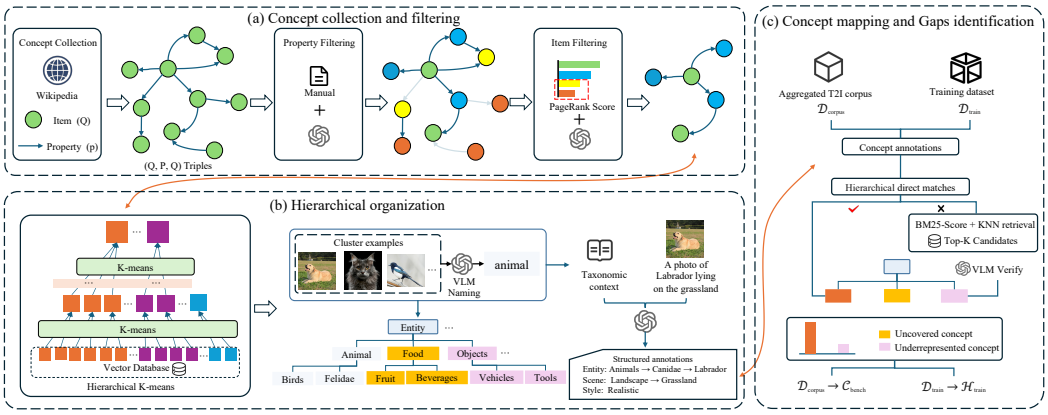


Figure 2: **Graph construction and concept coverage analysis.** (a) Filtering retains visually relevant concepts and factual relations. (b) Hierarchical k-means and VLM naming organize concepts. (c) Annotation linking supports separate audits of the aggregated corpus and each training dataset to select benchmark and supplementation targets, respectively.

relevance with a vision-language model (VLM). Its filtering prompt is refined through sample review and revision before application to the remaining concepts. Rejected concepts are revisited through retrieval to reduce false removals.

Hierarchical organization. To address uneven graph depth and semantic granularity, we apply hierarchical k-means (Vo et al., 2024), recursively clustering concept embeddings and then their centroids (Figure 2(b)). We select branch granularity and depth using cosine silhouette, ARI across repeated runs, local neighborhood consistency, and agreement between adjacent levels (Appendices C.1 and C.2). A VLM names each cluster from concepts near its centroid; the highest-level names are checked for semantic consistency. The hierarchy groups finer categories under entities, scenes, styles, and works, supporting coverage analysis, gap identification, and category priors for training.

3.2 T2I Corpus Construction and Concept Mapping

Aggregated T2I corpus. Our analysis corpus contains 370M annotated image–caption pairs from 20 public datasets and additional web and synthetic collections (Appendix Table 6). Each pair is one sample. We analyze this corpus and each target training dataset separately using the procedure in Figure 2(c).

Concept annotation and linking. For $\mathcal{D} = \{(I_i, t_i)\}_{i=1}^{N}$, Qwen3.5-9B (Qwen Team, 2026b) extracts entity, scene, and style annotations using the hierarchy as taxonomic context; existing annotations are retained. We resolve category labels from coarse to fine and match normalized names and aliases to canonical concepts, using categories to disambiguate matches. For unresolved annotations, we retrieve k nearest candidates from compatible categories and rank them by cosine similarity and normalized BM25 scores (Robertson & Zaragoza, 2009). A VLM verifies matches against the annotation context and candidate identities, resolving 93.11% of submitted queries.

Concept coverage statistics. For concept c in the canonical concept set $\mathcal{C}$, sample support is

$$n_c(\mathcal{D}) = \sum_{i=1}^{N} \mathbf{1}\left[c \in \widehat{\mathcal{C}}_i\right], \tag{1}$$

where $\widehat{\mathcal{C}}_i \subseteq \mathcal{C}$ contains the concepts linked to sample i. Each concept is counted at most once per sample, regardless of repeated mentions or multiple hierarchy memberships. Zero support means that the annotation and linking procedure did not observe the concept. Among observed concepts, $\mathcal{C}_{\text{corpus}} = \{c \in \mathcal{C} \mid n_c(\mathcal{D}_{\text{corpus}}) > 0\}$, the top 1% account for 94.98% of total sample support (Appendix Figure 5b). These counts identify unobserved and sparsely supported benchmark candidates.

3.3 IDENTIFYING CONCEPT COVERAGE GAPS

Coverage gap identification. Let $c(p)$ be the canonical concept at hierarchy position p, $\mathcal{C}(p)$ the set of canonical concepts at p or its descendants, and $\mathcal{B}(u)$ the eligible child positions of parent u. We define subtree support and a threshold based on nonzero sibling support as

$$S_p(\mathcal{D}) = \sum_{c \in \mathcal{C}(p)} n_c(\mathcal{D}), \qquad \tau_u(\mathcal{D}) = \alpha \,\mathrm{median}\{n_{c(r)}(\mathcal{D}) \mid r \in \mathcal{B}(u), n_{c(r)}(\mathcal{D}) > 0\}. \tag{2}$$

Both are computed separately for $\mathcal{D}_{\mathrm{corpus}}$ and $\mathcal{D}_{\mathrm{train}}$, counting canonical concepts without duplication across hierarchy memberships. We then distinguish absent subtree support from sparse support relative to siblings:

$$\mathrm{Uncovered}(p) = \mathbf{1}[S_p(\mathcal{D}) = 0], \qquad \mathrm{Underrep}(p \mid u) = \mathbf{1}\big[0 < n_{c(p)}(\mathcal{D}) < \tau_u(\mathcal{D})\big]. \tag{3}$$

The first criterion requires zero support for p and all its descendants; the second compares the target's canonical support with the local threshold. Evidence from other hierarchy memberships of the same concept is preserved. Appendix B.2 specifies eligibility for $\mathcal{B}(u)$ and the value of α.

Target selection. We merge candidates by canonical identifier, remove abstract labels and visually redundant variants, and screen for stable identity, domain relevance, and distinctive visual content. Applied separately to $\mathcal{D}_{\mathrm{corpus}}$ and $\mathcal{D}_{\mathrm{train}}$, this yields benchmark concepts $\mathcal{C}_{\mathrm{bench}} \subseteq \mathcal{C}$ and supplementation targets $\mathcal{H}_{\mathrm{train}} \subseteq \mathcal{C}$. Benchmark concepts also require usable reference evidence. In the audited corpus, 60.8% of these concepts are unobserved and 96.1% have at most ten supporting samples. Appendix D.1 gives their category distributions and counting conventions (Figure 10).

4 COVERAGE-GUIDED TRAINING DATA REFINEMENT AND WORLD KNOWLEDGE EVALUATION

We use $\mathcal{H}_{\mathrm{train}}$ for supplementary synthesis and the hierarchy for training-data sampling. KnowGap-Bench evaluates concept identity and composition separately on $\mathcal{C}_{\mathrm{bench}}$ (Section 3.3).

4.1 COVERAGE-GUIDED TRAINING DATA REFINEMENT

Visual roles and scene graph sampling. We synthesize examples for $\mathcal{H}_{\mathrm{train}}$ from scene graphs (Gao et al., 2026) $\mathcal{S} = (\mathcal{O}, \mathcal{A}, \mathcal{R}, \mathcal{Z})$, specifying objects, attributes, directed relations, and global scene or style conditions (Figure 3(a)). Concept identity, Wikidata type, and one-hop facts determine visual roles: entities are objects; scenes and styles are global conditions; works and activities appear through associated people, objects, actions, or environments. With targets fixed, we vary the numbers of objects, attributes, and relations, drawing auxiliary objects and global conditions from compatible hierarchy groups. One-hop facts constrain combinations; attribute and relation sampling respects object categories, participant types, and direction. Captions are screened for target identity, plausibility, and constraint preservation.

Caption refinement and data curation. A language model refines captions for fluency using the sampled graph and factual context. Each caption t is parsed back into $\widehat{\mathcal{S}}(t)$ by a language model. Agreement with the sampled graph is measured by tuple-level F1:

$$F_{\mathrm{tuple}}(\mathcal{S}, \widehat{\mathcal{S}}(t)) = \frac{2|\mathcal{T}(\mathcal{S}) \cap \mathcal{T}(\widehat{\mathcal{S}}(t))|}{|\mathcal{T}(\mathcal{S})| + |\mathcal{T}(\widehat{\mathcal{S}}(t))|}, \tag{4}$$

where $\mathcal{T}$ returns a multiset of normalized object, attribute, relation, and global-condition tuples. Matching preserves multiplicity, attribute bindings, and relation direction. Within each batch, we rank captions by F1 for each target and complexity level.

Validated captions specify composition, while curated reference images guide appearance during synthesis. VQAScore (Lin et al., 2024) selects aligned image–caption pairs, which retain concept annotations and form $\mathcal{D}_{\mathrm{aug}}$. We combine these with $\mathcal{D}_{\mathrm{train}}$ resampled by category priors over semantic groups defined by concept annotations and parent categories (Appendix F).

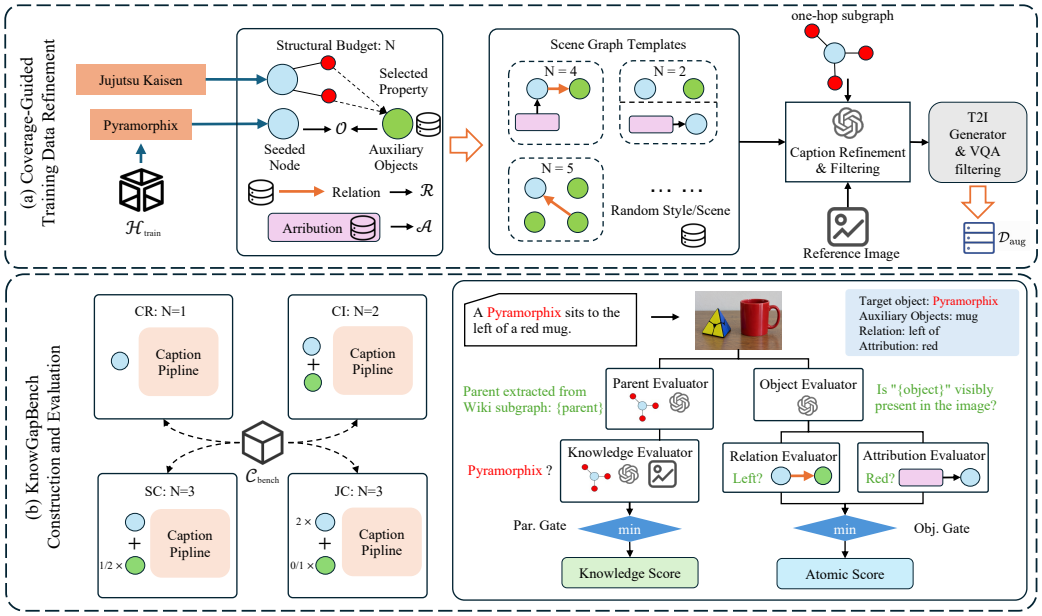


Figure 3: **Coverage-guided synthesis and KnowGapBench evaluation.** (a) Scene graphs and one-hop facts constrain captions, while reference images guide synthesis and alignment filtering selects supplementary pairs. (b) Four tasks evaluate identity with references (Knowledge) and composition from generated images alone (Atomic), with dependency gating.

4.2 KnowGapBench: Construction and Structured Evaluation

Benchmark construction. KnowGapBench contains 3,698 prompts covering 3,286 target concepts in $\mathcal{C}_{\text{bench}}$, selected from the corpus coverage gaps identified in Section 3.3. The concept inventory spans entities, scenes, styles, and works across 30 named hierarchical categories. Task counts and category distributions appear in Appendix Table 8 and Appendix D.1. Prompts use the scene graph and caption pipeline in Section 4.1 (Figure 3(b)); refinement preserves canonical names and adds no identifying details beyond the task specification.

Task design. For entity tasks, the structural budget $N = |\mathcal{O}| + |\mathcal{A}| + |\mathcal{R}|$ counts objects, attribute bindings, and directed relations, excluding global scene or style conditions. *Concept Realization* (CR; $N = 1$) tests one target with minimal additional constraints. *Contextual Integration* (CI; $N = 2$) adds an auxiliary object to test concept depiction in context. *Structured Composition* (SC; $N = 3$) tests object coexistence, attribute binding, or a directed relation. *Joint Concept Composition* (JC; $N = 3$) replaces an auxiliary object with a second target entity. SC and JC share configurations $(|\mathcal{O}|, |\mathcal{A}|, |\mathcal{R}|) \in \{(3, 0, 0), (2, 1, 0), (2, 0, 1)\}$. JC pairs are selected by greedy compatibility ranking without concept reuse across pairs. Other visual roles use their corresponding placements and templates.

Structured evaluation. *Knowledge* measures concept identity; *Atomic* measures composition. A multimodal evaluator rates concept identity from 0 to 4 using the generated image, canonical identity, Wikidata context, and audited references. Ratings range from incorrect or coarse category matches to clear visual matches. Reference images are withheld from generators. Atomic uses image-only binary judgments for specified auxiliary objects, attribute bindings, and directed relations.

Knowledge ratings are divided by four; Atomic judgments use yes=1 and no=0. These independently obtained raw scores in $[0, 1]$ are gated by their prerequisites (Cho et al., 2024):

$$s_i = \min\big(\{r_i\} \cup \{s_j : j \in D_i\}\big), \tag{5}$$

where r_i and s_i are raw and gated scores, and D_i indexes prerequisite judgments. Entity Knowledge depends on *Parent*, the presence of its graph-derived parent category. Attributes and relations depend

on participant presence: Parent for target entities and *Object Presence* for auxiliaries. Specific identity does not gate composition; an identity error alone therefore does not invalidate otherwise correct compositional judgments.

For each prompt, Knowledge averages the gated scores of its target concepts, and Atomic averages the gated judgments for its applicable compositional constraints. Atomic is omitted when no questions apply. We first average scores within each concept or JC pair, then take an equal-weight average across these units and scale the result to 0–100. Overall pools Knowledge records across tasks before this aggregation; it is not the unweighted mean of the four task scores.

5 Experiments

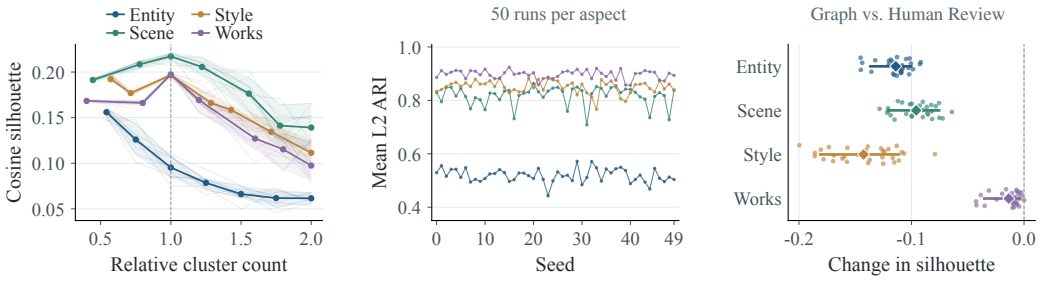
5.1 Experimental Setup

Evaluated models. We evaluate 13 T2I models (Table 2). Open-weight models are Boogu-Image-0.1-Base (Chen et al., 2026; Boogu Team, 2026), HunyuanImage-3.0-Instruct (Cao et al., 2025), Z-Image, Qwen-Image (Wu et al., 2025), GLM-Image (Z.ai, 2026), SenseNova-U1.5-8B-MoT (Diao et al., 2026), InternVL-U (Tian et al., 2026), and UniWorld-V1 fine-tuned on Fine-T2I (Lin et al., 2025). Hosted models are GPT-Image-2 (OpenAI, 2026), Gemini-3.1-Flash-Image-Preview (Google DeepMind, 2026), Seedream-5.0-Lite (ByteDance Seed, 2026), Qwen-Image-2.0-Pro (Zhao et al., 2026), and FLUX.2 [pro] (Black Forest Labs, 2025).

Training setting. For fine-tuning, we use the Fine-T2I dataset, which contains 6,177,306 image-text pairs. We train UniWorld-V1 both with and without CGTR. CGTR replaces approximately 5% of entries with supplementary pairs and resamples the remainder using category priors. Both runs use the same initial weights, generative objective, trainable parameters, and one epoch of equal size (Appendix F).

Evaluator choice and validation. We compare evaluators on 200 shared image–prompt pairs, 50 per task (Appendix Table 11). Annotators independently assign 0–4 identity ratings and binary atomic judgments, blinded to model identities and automatic scores. Before adjudication, we measure agreement using Spearman's ρ and linearly weighted Cohen's κ_w for Knowledge, and accuracy and Cohen's κ for atomic judgments. Qwen3.8-27B (Qwen Team, 2026a) provides consistent batch scoring with the references, Wikidata context, and dependency rules in Section 4.2.

5.2 Hierarchical Clustering and Graph Analysis



(a) Granularity under different cluster counts. **(b)** Stability across 50 random seeds. **(c)** Silhouette change after human review.

Figure 4: Hierarchy diagnostics for clustering granularity, initialization stability, and human-reviewed memberships.

We assess semantic coherence using cosine silhouette, neighbor precision, and triplet concordance. On a diagnostic set of 706 concepts, we vary the finest-level cluster count and compare L2 assignments across 50 random seeds using ARI. Reviewers independently assign categories to 738 concepts without seeing the original labels.

Appendix Table 7 shows higher silhouette (0.097 vs. -0.139) and neighbor precision at $k = 5$ (55.5% vs. 9.0%) than within-parent label permutations. Triplet concordance reaches 76.5% at L2 and 68.5%

at L1, indicating coherent local and intermediate semantic structure. Increasing the finest-level cluster count reduces separation (Figure 4(a)), suggesting that finer partitions are not always beneficial and supporting branch-specific granularity. Across 50 random initializations, mean L2 ARI is 0.519, 0.818, 0.850, and 0.897 for entities, scenes, styles, and works, respectively (Figure 4(b)). Entities show lower stability, consistent with the greater semantic diversity of this branch. Human review of boundary concepts produces similar branch-level silhouette patterns (Figure 4(c)), showing that semantic organization is largely preserved even for difficult cases near cluster boundaries.

The final graph contains 254,067 canonical concepts organized under four broad branches: entities, scenes, styles, and works. The hierarchy spans multiple semantic levels, from broad categories to individual concepts. Representative paths and the full hierarchy statistics are provided in Appendix C.1.

5.3 Evaluation on KnowGapBench

We evaluate 13 T2I models on the four KnowGapBench tasks. The completed evaluations use the frozen benchmark release, with one generated image per prompt when generation succeeds. Qwen3.8-27B scores all models using the protocol in Section 4.2. Means exclude unavailable scores but retain observed zeros, including those introduced by dependency gating. Scores are averaged within each concept and then across concepts with equal weight; JC uses concept pairs as the averaging units. Overall pools Knowledge scores across all four tasks using the same concept-macro aggregation.

Table 2: Results on KnowGapBench (higher is better). Par., Know., and Atom. denote Parent, Knowledge, and Atomic scores. Best results are bold with blue shading; second-best results are underlined with lighter shading.

Model	CR		CI		SC			JC			Overall
	Par.	Know.	Par.	Know.	Par.	Know.	Atom.	Par.	Know.	Atom.	Know.
Proprietary models											
GPT-Image-2	89.8	**61.9**	90.2	**59.6**	88.8	**50.7**	**63.0**	82.4	**44.5**	**66.5**	**53.9**
Gemini 3.1 Flash	91.5	57.5	89.9	53.7	89.1	42.4	50.5	82.5	42.2	58.1	49.6
Seedream 5.0	91.4	46.5	**92.7**	41.4	**90.4**	32.1	52.3	**83.4**	32.9	62.1	38.8
FLUX.2 Pro	**92.0**	42.7	90.8	35.7	89.8	27.8	46.2	82.7	29.6	58.0	35.1
Qwen-Image-2.0-Pro	91.3	40.6	90.6	35.2	88.5	26.5	47.1	79.6	28.0	55.4	33.9
Open-weight models											
HunyuanImage-3.0	89.2	34.9	87.7	29.7	87.5	24.0	41.1	73.8	22.4	48.1	28.1
Boogu-Image	87.6	29.8	83.8	24.8	83.4	19.0	39.2	66.3	19.0	43.4	24.3
SenseNova-U1.5	86.6	27.8	84.9	23.6	83.0	17.9	37.6	70.1	17.3	45.6	22.8
Z-Image	83.5	26.8	79.7	24.6	76.1	19.0	35.0	58.5	14.0	38.1	21.6
Qwen-Image	84.4	27.5	82.2	24.0	80.4	17.7	38.9	60.6	16.1	41.8	21.5
UniWorld Fine-T2I	78.2	19.3	70.3	15.7	67.5	12.6	25.8	47.5	11.0	27.6	15.3
GLM-Image	76.5	20.9	65.5	15.5	67.6	12.8	26.4	42.4	11.2	27.2	15.3
InternVL-U	77.2	14.7	67.4	11.6	69.2	7.6	24.6	46.0	6.4	28.2	10.4

Concept identity. GPT-Image-2 obtains the highest Knowledge score in all four tasks, from 61.9 in CR to 44.5 in JC (Table 2). HunyuanImage-3.0 is the strongest model evaluated from local weights, scoring 34.9 and 22.4 in these two settings. Recognizing a broad category is easier than recovering the specific concept: on the same entity subset of CR, GPT-Image-2 scores 89.8 for Parent but 56.6 for Knowledge. Each task characterizes a distinct target set; task-level scores therefore reflect both concept identity and compositional requirements.

Composition. Table 3 separates object presence, attribute binding, and relation compliance. Atomic components are recomputed over the pooled SC and JC samples. GPT-Image-2 leads all three components, with scores of 68.0 for objects, 50.1 for attributes, and 61.0 for relations. Across models, attribute scores are lower than object scores on their respective prompt subsets, identifying attribute binding as a distinct evaluation axis.

The visual-preference ranking is complementary: GLM-Image achieves the second-highest HPSv3 (Ma et al., 2025) score, while its SC Knowledge score is 12.8. Reporting these dimensions separately distinguishes visual appeal from concept fidelity. Style evaluates compliance with the global style condition, while HPSv3 measures visual preference on its native scale.

Semantic categories. The detailed breakdown is in Appendix Table 12; GPT-Image-2 leads nine of the ten semantic categories. Gemini 3.1 Flash leads biological kinds with 44.9, compared with 39.3 for GPT-Image-2; leading scores exceed 70 for vehicles and devices and for activities and events. This category-level analysis identifies where targeted concept supplementation can be most informative.

The breakdown covers 2,086 individual targets in CR, CI, and SC; the 600 JC pairs are reported separately. These ten evaluation categories differ from the four top-level branches of the concept hierarchy. In JC, Knowledge averages the two target scores, whereas Parent is the minimum of the available target Parent scores. JC Knowledge therefore measures average target fidelity, not simultaneous success on both identities.

Table 3: Composition, style, and visual preference. Obj., Attr., and Rel. denote objects, attributes, and relations. Highlighting follows Table 2.

Model	Composition			Style	Preference
	Obj.	Attr.	Rel.	Style	HPSv3
Proprietary models					
GPT-Image-2	**68.0**	**50.1**	**61.0**	**80.8**	11.9090
Gemini 3.1 Flash	55.5	40.4	52.9	76.3	**12.8017**
Seedream 5.0	56.1	42.9	60.8	71.6	10.6754
FLUX.2 Pro	51.0	40.7	51.6	69.6	10.8213
Qwen-Image-2.0-Pro	50.3	43.3	50.8	68.2	11.1057
Open-weight models					
HunyuanImage-3.0	43.9	37.3	41.5	65.3	10.6191
Boogu-Image	42.7	32.0	36.3	74.3	11.1790
SenseNova-U1.5	41.0	33.7	37.5	68.3	10.8016
Z-Image	39.5	29.5	25.1	69.5	10.2169
Qwen-Image	42.6	33.3	32.3	63.7	10.4342
UniWorld Fine-T2I	29.0	26.1	14.6	59.1	10.6176
GLM-Image	29.6	21.4	18.3	50.9	12.3503
InternVL-U	26.8	24.3	18.6	60.1	9.6942

5.4 Effectiveness of CGTR

CommonConceptBench evaluates familiar concepts disjoint from the supplementation targets. It uses the same graph-based construction as KnowGapBench, with 1,000 prompts covering 632 concepts across CR, CI, and SC (Appendix F.2).

With CGTR, UniWorld-V1's overall Knowledge rises from 15.3 to 22.5 on KnowGapBench (+7.2 points) and from 62.1 to 65.9 on CommonConceptBench (Table 4). CommonConceptBench gains are larger in CI and SC than in CR. Parent and Atomic distinguish concept gains from compositional effects.

Table 4: Effect of CGTR on UniWorld-V1 trained on Fine-T2I. Subscripts in Overall indicate score-point gains relative to the baseline. Dashes denote the absent JC task in CommonConceptBench.

Setting	CR		CI			SC			JC			Overall
	Par.	Know.	Par.	Know.	Atom.	Par.	Know.	Atom.	Par.	Know.	Atom.	Know.
KnowGapBench												
Baseline	78.2	19.3	70.3	15.7	—	67.5	12.6	25.8	47.5	11.0	27.6	15.3
CGTR	90.1	32.3	80.6	24.2	—	73.4	17.9	24.6	55.7	15.4	33.8	**22.5** $_{\uparrow 7.2}$
CommonConceptBench												
Baseline	94.0	88.9	82.1	46.8	12.9	78.5	47.1	47.7	—	—	—	62.1
CGTR	96.3	90.8	84.0	51.2	20.0	81.0	51.7	49.2	—	—	—	**65.9** $_{\uparrow 3.8}$

6 Conclusion

We presented CGTR for auditing concept coverage and refining T2I training data. A hierarchical world knowledge graph maps heterogeneous annotations to canonical concepts, enabling coverage measurement, data tracing, and concept-aware sampling. Auditing 370M records reveals substantial gaps and concentration in concept support. These measurements guide KnowGapBench, which separates concept identity from composition and shows that broad-category recognition does not ensure concept fidelity. CGTR audits each target dataset independently, synthesizes examples for insufficiently supported concepts, and resamples the original data using category priors. Refining Fine-T2I raises UniWorld-V1's overall Knowledge from 15.3 to 22.5, preserving performance on familiar concepts outside supplementation targets. These results support concept-level coverage as a basis for targeted improvements to training-data composition.

AI USE STATEMENT

Generative AI tools were used in this work to assist with software implementation and debugging, and to improve the clarity, grammar, and presentation of the manuscript. AI models were also used as automated evaluators in the experimental pipeline; their roles, inputs, scoring protocols, and validation procedures are described in Sections 4.2 and 5.1 and Appendix E. All AI-assisted code, text, and experimental outputs were reviewed and verified by the authors. The authors made the final decisions regarding the methodology, analysis, and conclusions, and take full responsibility for the content of this work.

ETHICS STATEMENT

This work studies concept coverage, evaluation, and data refinement for text-to-image models. The experiments use existing image–text datasets, web-collected or synthetic data, and model-generated images for research purposes. Human review is limited to semantic annotation and evaluation judgments and does not require collecting sensitive personal information. We use the collected resources only for research analysis and evaluation and follow the usage conditions of the corresponding data sources where applicable.

REPRODUCIBILITY STATEMENT

We provide the details required to reproduce our experiments in the main paper and supplementary material. Sections 3–4 and Appendices A–E document graph construction, concept mapping, coverage analysis, hierarchy diagnostics, KnowGapBench construction, evaluator prompts, and score aggregation. Section 5.1 and Appendix F specify the training setup, sampling strategy, optimization settings, random seeds, and evaluation protocol. Upon publication, we will release the CGTR construction pipeline and KnowGapBench, together with the benchmark annotations, prompts, evaluation code, and experiment configurations, subject to the licenses and usage conditions of the underlying data sources.

REFERENCES

Mohammad Kalim Akram, Saba Sturua, Nastia Havriushenko, Quentin Herreros, Michael Günther, Maximilian Werk, and Han Xiao. jina-embeddings-v5-text: Task-targeted embedding distillation. *arXiv preprint arXiv:2602.15547*, 2026. URL https://arxiv.org/abs/2602.15547.

David Beniaguev. Synthetic faces high quality - text 2 image (SFHQ-T2I) dataset, 2024. URL https://github.com/SelfishGene/SFHQ-T2I-dataset.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence, November 2025. URL https://bfl.ai/blog/flux-2. Technical blog. Accessed: 2026-09-25.

Boogu Team. Boogu-Image-0.1-Base (revision fa78121), 2026. URL https://huggingface.co/Boogu/Boogu-Image-0.1-Base/tree/fa78121.

Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117, 1998. doi: 10.1016/S0169-7552(98)00110-X. URL https://doi.org/10.1016/S0169-7552(98)00110-X.

ByteDance Seed. Seedream 5.0 Lite, 2026. URL https://seed.bytedance.com/en/seedream5_0_lite. Official model page. Accessed: 2026-09-25.

Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, et al. HunyuanImage 3.0 technical report. *arXiv preprint arXiv:2509.23951*, 2025. doi: 10.48550/arXiv.2509.23951. URL https://arxiv.org/abs/2509.23951.

Guoxuan Chen, Chufeng Xiao, Haoran Yang, Siyue Xie, Binxiao Huang, Ming Zhang, Cheuk Him Chau, Xinyu Fu, Yingzhao Lian, Tom S. Y. Li, Jintao Lin, Bowen Dong, Zian Qian, Yuhao Liu, Yuxuan Hu, Weikang Shi, Bin Zou, Bowen Zheng, Haoxuan Che, Chang Chen, Yuyang He, Heyang Sun, Tianyu Huang, Chong Hou Choi, Cheng Gong, Han Shi, Haoli Bai, Xihui

Liu, Hongsheng Li, Qifeng Chen, Chao Huang, Rui Liu, and Chenyang Lei. Boogu-Image-0.1: Boosting open agentic multimodal generation via understanding under a minimal budget, 2026. URL https://arxiv.org/abs/2607.13125.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. BLIP3-o: A family of fully open unified multimodal models–architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025a. URL https://arxiv.org/abs/2505.09568.

Junying Chen, Zhenyang Cai, Pengcheng Chen, Shunian Chen, Ke Ji, Xidong Wang, Yunjin Yang, and Benyou Wang. ShareGPT-4o-Image: Aligning multimodal models with GPT-4o-level image generation, 2025b. URL https://arxiv.org/abs/2506.18095.

Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W. Cohen. Re-Imagen: Retrieval-Augmented Text-to-Image Generator. *arXiv preprint arXiv:2209.14491*, 2022. URL https://arxiv.org/abs/2209.14491.

Zhihong Chen, Xuehai Bai, Yang Shi, Chaoyou Fu, Huanyu Zhang, Haotian Wang, Xiaoyan Sun, Zhang Zhang, Liang Wang, Yuanxing Zhang, Pengfei Wan, and Yi-Fan Zhang. OpenGPT-4o-Image: A comprehensive dataset for advanced image generation and editing, 2025c. URL https://arxiv.org/abs/2509.24900.

Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.18235.

Dawei Dai, Long Xu, Yutang Li, Yuanhui Zhang, and Shuyin Xia. Humanvlm: Foundation for human-scene vision-language model. *Information Fusion*, 123:103271, 2025. doi: 10.1016/j.inffus.2025.103271. URL https://doi.org/10.1016/j.inffus.2025.103271.

Haiwen Diao, Jiahao Wang, Chenjing Ding, Hanming Deng, Jiangnan Chen, Ruixi Zhang, Ruohui Wang, Wenwen Tong, Xiangyu Fan, Yubo Wang, et al. SenseNova-U1.5: Towards native unified visual intelligence. *arXiv preprint arXiv:2609.11929*, 2026. URL https://arxiv.org/abs/2609.11929.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander T Toshev, and Vaishaal Shankar. Data filtering networks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KAk6ngZ09F.

Rongyao Fang, Aldrich Yu, Chengqi Duan, Linjiang Huang, Shuai Bai, Yuxuan Cai, Kun Wang, Si Liu, Xihui Liu, and Hongsheng Li. FLUX-reason-6m & PRISM-bench: A million-scale text-to-image reasoning dataset and comprehensive benchmark. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=cPzgZnpVbN.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah M Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=dVaWCDMBof.

Yifan Gao, Jinpeng Lin, Min Zhou, Chuanbin Liu, Hongtao Xie, Tiezheng Ge, and Yuning Jiang. TextPainter: Multimodal text image generation with visual-harmony and text-comprehension for poster design. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 7236–7246. ACM, 2023. doi: 10.1145/3581783.3611929. URL https://doi.org/10.1145/3581783.3611929.

Ziqi Gao, Weikai Huang, Jieyu Zhang, Aniruddha Kembhavi, and Ranjay Krishna. Generate any scene: Scene graph driven data synthesis for visual generation training. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=EwdWR6lfvW.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=Wbr51vK331.

Google DeepMind. Gemini 3.1 Flash Image, 2026. URL https://deepmind.google/models/model-cards/gemini-3-1-flash-image/. Model card, published February 26, 2026. Accessed: 2026-09-25.

Shuhao Han, Haotian Fan, Jiachen Fu, Liang Li, Tao Li, Junhui Cui, Yunqiu Wang, Yang Tai, Jingwei Sun, Chun-Le Guo, and Chongyi Li. Evalmuse-40k: A fine-grained benchmark with comprehensive human annotations for text-to-image generation model alignment evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4583–4591, 2026. doi: 10.1609/aaai.v40i6.42458. URL https://doi.org/10.1609/aaai.v40i6.42458.

Mengchao He, Yuliang Liu, Zhibo Yang, Sheng Zhang, Canjie Luo, Feiyu Gao, Qi Zheng, Yongpan Wang, Xin Zhang, and Lianwen Jin. ICPR2018 contest on robust reading for multi-type web images. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 7–12. IEEE, 2018. doi: 10.1109/ICPR.2018.8546143. URL https://doi.org/10.1109/ICPR.2018.8546143.

Hsin-Ping Huang, Xinyi Wang, Yonatan Bitton, Hagai Taitelbaum, Gaurav Singh Tomar, Ming-Wei Chang, Xuhui Jia, Kelvin C.K. Chan, Hexiang Hu, Yu-Chuan Su, and Ming-Hsuan Yang. KITTEN: A knowledge-integrated evaluation of image generation on visual entities. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=wejaKS9Ps0.

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench++: An Enhanced and Comprehensive Benchmark for Compositional Text-to-image Generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3563–3579, 2025a. doi: 10.1109/TPAMI.2025.3531907. URL https://doi.org/10.1109/TPAMI.2025.3531907.

Ziwei Huang, Wanggui He, Quanyu Long, Yandi Wang, Haoyuan Li, Zhelun Yu, Fangxun Shu, Weilong Dai, Hao Jiang, Fei Wu, and Leilei Gan. T2I-FactualBench: Benchmarking the factuality of text-to-image models with knowledge-intensive concepts. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27501–27524, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1334. URL https://aclanthology.org/2025.acl-long.1334/.

Nupur Kumari, Xi Yin, Jun-Yan Zhu, Ishan Misra, and Samaneh Azadi. Generating multi-image synthetic data for text-to-image customization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16524–16534, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Kumari_Generating_Multi-Image_Synthetic_Data_for_Text-to-Image_Customization_ICCV_2025_paper.html.

Hui Li, Mingwang Xu, Yun Zhan, Shan Mu, Jiaye Li, Kaihui Cheng, Yuxuan Chen, Tan Chen, Mao Ye, Jingdong Wang, and Siyu Zhu. OpenHumanVid: A large-scale high-quality dataset for enhancing human-centric video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7752–7762, June 2025a. URL https://openaccess.thecvf.com/content/CVPR2025/html/Li_OpenHumanVid_A_Large-Scale_High-Quality_Dataset_for_Enhancing_Human-Centric_Video_Generation_CVPR_2025_paper.html.

Jialuo Li, Wenhao Chai, Xingyu Fu, Haiyang Xu, and Saining Xie. Science-t2i: Addressing scientific illusions in image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2734–2744, June 2025b. URL https://openaccess.thecvf.com/content/CVPR2025/html/Li_Science-T2I_Addressing_Scientific_Illusions_in_Image_Synthesis_CVPR_2025_paper.html.

Xianhang Li, Haoqin Tu, Mude Hui, Zeyu Wang, Bingchen Zhao, Junfei Xiao, Sucheng Ren, Jieru Mei, Qing Liu, Huangjie Zheng, Yuyin Zhou, and Cihang Xie. What if we recaption billions of web images with LLaMA-3? In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 35957–35976. PMLR, 13–19 Jul 2025c. URL https://proceedings.mlr.press/v267/li25ch.html.

Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld-v1: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025. URL https://arxiv.org/abs/2506.03147.

Jinpeng Lin, Min Zhou, Ye Ma, Yifan Gao, Chenxi Fei, Yangjian Chen, Zhang Yu, and Tiezheng Ge. AutoPoster: A highly automatic and content-aware design system for advertising poster generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 1250–1260. ACM, 2023. doi: 10.1145/3581783.3611930. URL https://doi.org/10.1145/3581783.3611930.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *Computer Vision – ECCV 2024*, pp. 366–384, 2024. doi: 10.1007/978-3-031-72673-6_20. URL https://arxiv.org/abs/2404.01291.

Xu Ma, Yitian Zhang, Qihua Dong, and Yun Fu. Fine-T2I: An open, large-scale, and diverse dataset for high-quality T2I fine-tuning, 2026. URL https://arxiv.org/abs/2602.09439.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. HPSv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15086–15095, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Ma_HPSv3_Towards_Wide-Spectrum_Human_Preference_Score_ICCV_2025_paper.html.

Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Fanqing Meng, Kun-Peng Ning, Bin Zhu, and Li Yuan. WISE: World knowledge-informed semantic evaluation for text-to-image generation. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=f1q8dFnxtY.

OpenAI. GPT Image 2, 2026. URL https://developers.openai.com/api/docs/models/gpt-image-2. OpenAI API model documentation. Accessed: 2026-09-25.

Shintaro Ozaki, Tomoyuki Jinno, Kazuki Hayashi, Yusuke Sakai, Jingun Kwon, Hidetaka Kamigaito, Katsuhiko Hayashi, Manabu Okumura, and Taro Watanabe. TextTIGER: Text-based Intelligent Generation with Entity Prompt Refinement for Text-to-Image Generation. *arXiv preprint arXiv:2504.18269*, 2025. URL https://arxiv.org/abs/2504.18269.

Shubham Parashar, Zhiqiu Lin, Tian Liu, Xiangjue Dong, Yanan Li, Deva Ramanan, James Caverlee, and Shu Kong. The neglected tails in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12988–12997, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Parashar_The_Neglected_Tails_in_Vision-Language_Models_CVPR_2024_paper.html.

Qwen Team. Qwen3.8-27B, 2026a. URL https://huggingface.co/Qwen/Qwen3.8-27B. Official model card. Accessed: 2026-09-25.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026b. URL https://qwen.ai/blog?id=qwen3.5.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019. URL https://doi.org/10.1561/1500000019.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. URL https://arxiv.org/abs/2210.08402.

Hengyu Shen, Tiancheng Gu, Bin Qin, Lan Wu, Yuling Wu, Shuo Tan, Zelong Sun, Jun Wang, Nan Wu, Xiang An, Weidong Cai, Ziyong Feng, and Kaicheng Yang. DanQing: An up-to-date large-scale chinese vision-language pre-training dataset, 2026. URL https://arxiv.org/abs/2601.10305.

Baoguang Shi, Cong Yao, Minghui Liao, Mingkun Yang, Pei Xu, Linyan Cui, Serge Belongie, Shijian Lu, and Xiang Bai. ICDAR2017 competition on reading chinese text in the wild (RCTW-17). In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1429–1434. IEEE, 2017. doi: 10.1109/ICDAR.2017.233. URL https://doi.org/10.1109/ICDAR.2017.233.

Moo Hyun Son, Jintaek Oh, Sun Bin Mun, Jaechul Roh, and Sehyun Choi. World-To-Image: Grounding Text-to-Image Generation with Agent-Driven World Knowledge. *arXiv preprint arXiv:2510.04201*, 2025. URL https://arxiv.org/abs/2510.04201.

Jiao Sun, Deqing Fu, Yushi Hu, Su Wang, Royi Rassin, Da-Cheng Juan, Dana Alon, Charles Herrmann, Sjoerd van Steenkiste, Ranjay Krishna, and Cyrus Rashtchian. DreamSync: Aligning text-to-image generation with image understanding feedback. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5920–5945, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.304. URL https://aclanthology.org/2025.naacl-long.304/.

Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, et al. Journeydb: A benchmark for generative image understanding. *Advances in neural information processing systems*, 36:49659–49678, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/9bc59aff4685e39e1a8175d5303248a1-Abstract-Datasets_and_Benchmarks.html.

Haoyi Tao, Chaozheng Huang, Nan Wang, Han Lyu, Linfeng Zhang, Guolin Ke, and Xi Fang. Omniscience: A large-scale multi-modal dataset for scientific image understanding. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 9860–9870, 2026. URL https://huggingface.co/datasets/UniParser/OmniScience#citation.

Changyao Tian, Danni Yang, Guanzhou Chen, Erfei Cui, Zhaokai Wang, Yuchen Duan, Penghao Yin, Sitao Chen, Ganlin Yang, Mingxin Liu, et al. Internvl-u: Democratizing unified multimodal models for understanding, reasoning, generation and editing. *arXiv preprint arXiv:2603.09877*, 2026. URL https://arxiv.org/abs/2603.09877.

Kavana Venkatesh, Yusuf Dalva, Ismini Lourentzou, and Pinar Yanardag. RAVEL: Rare Concept Generation and Editing via Graph-driven Relational Guidance. *arXiv preprint arXiv:2412.09614*, 2024. URL https://arxiv.org/abs/2412.09614.

Huy V. Vo, Vasil Khalidov, Timothée Darcet, Théo Moutakanni, Nikita Smetanin, Marc Szafraniec, Hugo Touvron, camille couprie, Maxime Oquab, Armand Joulin, Herve Jegou, Patrick Labatut, and Piotr Bojanowski. Automatic data curation for self-supervised learning: A clustering-based

approach. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=G7p8djzWOl.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-Image Technical Report. *arXiv preprint arXiv:2508.02324*, 2025. URL https://arxiv.org/abs/2508.02324.

Hu Xu, Saining Xie, Xiaoqing Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying CLIP data. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5BCFlnfE1g.

William Yang, Xindi Wu, Zhiwei Deng, Esin Tureci, and Olga Russakovsky. Beyond objects: contextual synthetic data generation for fine-grained classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22498–22508, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Yang_Beyond_Objects_Contextual_Synthetic_Data_Generation_for_Fine-Grained_Classification_CVPR_2026_paper.html.

Junyan Ye, Dongzhi Jiang, Zihao Wang, Leqi Zhu, Zhenghao Hu, Zilong Huang, Jun He, Zhiyuan Yan, Jinghua Yu, Hongsheng Li, Conghui He, and Weijia Li. Echo-4o: Harnessing the Power of GPT-4o Synthetic Images for Improved Image Generation. *arXiv preprint arXiv:2508.09987*, 2025. URL https://arxiv.org/abs/2508.09987.

Z-Image Team. Z-Image: An Efficient Image Generation Foundation Model with Single-Stream Diffusion Transformer. *arXiv preprint arXiv:2511.22699*, 2025. URL https://arxiv.org/abs/2511.22699.

Z.ai. Glm-image: Auto-regressive for dense-knowledge and high-fidelity image generation. https://github.com/zai-org/GLM-Image, 2026. Accessed: 2026.

Daoan Zhang, Che Jiang, Ruoshi Xu, Biaoxiang Chen, Zijian Jin, Yutian Lu, Jianguo Zhang, Liang Yong, Jiebo Luo, and Shengda Luo. WorldGenBench: A World-Knowledge-Integrated Benchmark for Reasoning-Driven Text-to-Image Generation. *arXiv preprint arXiv:2505.01490*, 2025a. URL https://arxiv.org/abs/2505.01490.

Wenchao Zhang, Jiahe Tian, Runze He, Jizhong Han, Jiao Dai, Miaomiao Feng, Wei Mi, and Xiaodan Zhang. Align Beyond Prompts: Evaluating World Knowledge Alignment in Text-to-Image Generation. *arXiv preprint arXiv:2505.18730*, 2025b. URL https://arxiv.org/abs/2505.18730.

Bing Zhao, Chenfei Wu, Deqing Li, Hao Meng, Jiahao Li, Jie Zhang, et al. Qwen-Image-2.0 Technical Report. *arXiv preprint arXiv:2605.10730*, 2026. URL https://arxiv.org/abs/2605.10730.

Supplementary Material

From Knowledge Gaps to Better Generation: Auditing and Completing World Knowledge in Text-to-Image Training Data

This appendix documents graph context, dataset mapping, hierarchy diagnostics, KnowGapBench construction and evaluation, and coverage-guided training data refinement (CGTR). The mapping and evaluator prompts accompany their respective procedures.

Contents

	Appendix	Page
A	Knowledge Graph Filtering and Visual Context	17
B	Dataset Mapping and Coverage Analysis	18
	Inventory and coverage settings	
	Mapping prompt	19
C	Hierarchy Analysis	29
D	KnowGapBench Construction	33
	Task distributions and reference-image examples	
E	Benchmark Evaluation	35
	Protocol, visual relations, evaluator prompts, and results	
F	CGTR Training and Evaluation	42

A KNOWLEDGE GRAPH FILTERING AND VISUAL CONTEXT

Concept filtering retains stable visual identities with distinctive appearance or reusable domain value, while removing mapping noise and visually redundant or weakly grounded concepts. Table 5 summarizes the predicate groups used to extract attributes and one-hop context for caption construction and identity disambiguation.

Table 5: Predicate groups for visual attribute and context extraction. Property identifiers follow Wikidata; groups describe their operational use, and can overlap.

Use	**Group**	**Properties**
Appearance	Color	P462, P11220, P2827, P1340, P1884, P787, P534
	Material / ingredients	P186, P4543
	Shape / structure	P1419, P13768, P3739, P784
	State / orientation	P515, P3081, P7469
	Texture / foliage	P10906
	Other visual attributes	P8839, P8852, P6022, P3828, P4743, P2012, P8431, P149, P12516, P2739, P9302, P163, P237, P418, P4185, P6875, P5961
Context	Type, genre, basis, topic	P31, P279, P136, P144, P1269, P921
	Location / environment	P17, P495, P276, P706, P206, P669, P3032, P11633, P47, P3173, P4552, P4688, P403, P469, P4792, P974, P5998, P59, P2974
	Activities / participants	P641, P1344, P175, P161, P664, P710
	Parts / containment	P527, P1811, P361, P2670, P4330, P12443, P8308, P9813
	Works / narrative	P674, P1441, P840, P8411, P1080, P5800, P453
	Depiction, clothing, tools	P1303, P180, P3828, P7376, P13633, P3091, P1429, P822, P520

One-hop context preserves readable neighbor names and relation direction, prioritizing visually informative links such as characters, depicted subjects, participants, locations, and parts. The evaluator uses these facts to resolve the target identity; a fact contributes positive evidence only when it is consistent with visible image content.

B Dataset Mapping and Coverage Analysis

B.1 Dataset Inventory

Table 6 lists the processed source subsets and the samples available for annotation and coverage analysis. The main-text coverage statistics use the fixed snapshot in the coverage column.

Collected web images is an aggregate label for sources without a confirmed named public release; it also includes internally synthesized and rendered images. Known alternative caption variants and named T2I benchmark derivatives are excluded from the analysis corpus. Differences between the two columns should not be interpreted as linking failures. Public source names identify processed subsets and do not imply unrestricted image licensing. Sample counts are not adjusted for image overlap across sources.

Table 6: Dataset inventory. Coverage samples belong to the fixed mapping snapshot used in the analysis. Annotated samples have available tagging outputs; the two columns may refer to different snapshots. Counts describe the processed subsets, not the size of each public release.

Dataset	Coverage samples	Annotated samples
Collected web images	222,996,561	223,071,091
LAION-5B (Schuhmann et al., 2022)	100,770,722	100,772,414
DanQing100M (Shen et al., 2026)	13,988,084	13,988,943
BLIP3o-Pretrain (Chen et al., 2025a)	7,253,264	7,253,682
Fine-T2I (Ma et al., 2026)	6,177,306	6,177,306
HumanCaption-10M (Dai et al., 2025)	4,280,925	4,287,393
FLUX-Reason-6M (Fang et al., 2026)	3,803,352	3,798,079
JourneyDB (Sun et al., 2023)	3,603,142	3,603,192
OpenHumanVid (Li et al., 2025a)	2,856,067	2,856,089
OmniScience (Tao et al., 2026)	1,199,113	1,199,113
OSP1024-286k (Lin et al., 2025)	204,588	204,588
HumanCaption-HQ-311K (Dai et al., 2025)	201,247	201,354
SFHQ-T2I (Beniaguev, 2024)	121,987	121,987
Echo-4o-Image (Ye et al., 2025)	95,330	95,330
POSTER-TEXT (Gao et al., 2023)	65,779	65,779
BLIP3o-60k (Chen et al., 2025a; Lin et al., 2025)	51,433	58,859
AutoPoster (Lin et al., 2023)	38,570	38,570
ShareGPT-4o-Image (Chen et al., 2025b)	29,694	29,694
OpenGPT-4o-Image (Chen et al., 2025c)	29,132	29,132
RCTW-17 (Shi et al., 2017)	2,453	2,453
MTWI (He et al., 2018)	2,002	2,002
Aggregated corpus total	367,770,751	—

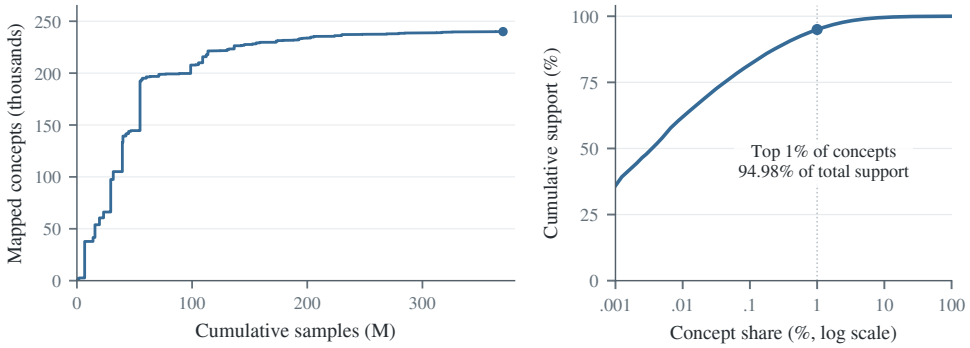
B.2 Coverage Audit Settings

The sibling comparison uses $\alpha = 0.1$ and requires at least eight eligible child positions with a total direct query frequency of at least 50. The median is computed over positive direct frequencies only. These settings apply to the shared gap identification procedure, with frequencies recalculated for each dataset. The uncovered criterion does not require an eligible sibling group. Ancestor accumulation counts each matched hierarchy position once per distinct ancestor; concept frequencies are computed before links are assigned to hierarchy positions.

The reported coverage snapshot contains 1,102,570,363 eligible query occurrences, of which 1,026,603,265 have accepted links. At least one link is present in 367,631,540 of the 367,770,751 analyzed samples. Of the 241,295 linked canonical concepts, 237,126 have a recoverable hierarchy category. These denominators distinguish query linking, sample linking, and hierarchy assignment.

Figure 5a tracks concept accumulation using 105 disjoint batches from the saved mapping index and 12 random batch orders with seed 50. Every plotted vertex is an exact union of the linked canonical identifiers in the batches added so far. The first sampled order is highlighted, and lines connect the measured vertices. The resulting trajectories quantify order dependence within this fixed corpus.

For the frequency concentration analysis in Figure 5b, observed concepts are sorted by sample support. The highest-frequency $\lceil 0.01|\mathcal{C}_{\text{corpus}}| \rceil = 2,413$ concepts contribute 94.98% of the total support of 973,079,372. A sample may support several concepts, so these proportions measure summed concept support rather than the fraction of samples containing a head concept.



(a) Concept accumulation. **(b)** Frequency concentration.

Figure 5: Coverage diagnostics for the 367.77M-sample corpus. (a) Distinct linked concepts across 12 random batch orders; the first order is highlighted. (b) Cumulative concept support ranked by frequency.

B.3 Hierarchical Annotation Prompt

Prompt extracts entity, scene, and style annotations for the concept-linking procedure in Section 3.2. We reproduce the original prompt used for tagging, including JSON keys and label names, for readability.

Mapping prompt

```
You are a hierarchical visual annotation assistant for multimodal
training datasets.

Task:
Using only the input image, generate hierarchical labels for three
aspects:  entities, scenes, and styles.

Input:
- image

Final output:
- Output valid JSON only.
- Do not output explanations, analysis, Markdown, comments, or extra
text.
- Use English labels.
- English IDs are for external mapping only; do not output English IDs
in the JSON.
- Select primary labels strictly from the taxonomy.
- Prefer secondary labels from the taxonomy.  If the list is clearly
unsuitable, a new stable, general, reusable English secondary label is
allowed.
- When uncertain, prefer a broader label or an empty value; do not
guess.

1. Output format

{
  "entities":  [
```

Mapping prompt (continued)

```
    {
      "primary_label":  "",
      "secondary_label":  "",
      "tertiary_label":  "",
      "factual_label":  ""
    }
  ],
  "scenes":  [
    {
      "primary_label":  "",
      "secondary_label":  "",
      "factual_label":  ""
    }
  ],
  "styles":  [
    {
      "primary_label":  "",
      "secondary_label":  ""
    }
  ]
}
Field requirements:
1.  The entity list may be []; output one item by default, at most
three.
2.  The scene list may be []; output one item by default.
3.  The style list may be []; output one item by default.
4.  For a nonempty entity list, each item must have nonempty primary,
secondary, and tertiary labels.  The tertiary label must not be
identical to its primary or secondary label.
5.  Do not output tertiary labels for scenes or styles.
6.  The factual label must be "" when reliable evidence is
unavailable.
7.  Do not output null, unknown, none, other, or multiple parallel
terms.  Do not mix languages except for taxonomy entries or
conventional fixed expressions.
```

2. Taxonomy

```
Notes:
1.  Primary labels form a closed set; use the names below exactly.
2.  Secondary lists are preferred candidates, not closed sets.  A
stable English secondary category may be added when the list does not
cover the image and sufficient evidence is available.
3.  Do not use specific instances, proper names, brands, works, place
names, model names, or animal breeds as secondary labels.
[Entities:  content subjects]
Portraits:
[individual portrait, body part, group portrait, celebrity,
occupation]
```

Mapping prompt (continued)

```
Characters:
[film or television character, game character, animation or comic
character, virtual persona, popular IP]
Activities:
[sport, festival, performance, event, meeting, award, gathering,
shopping, cooking]
Clothing:
[jewelry, footwear, top, skirt or dress, hosiery, hat, traditional
ethnic clothing, formal wear, uniform, swimwear, underwear]
Animals:
[birds, fish, reptiles, amphibians, insects, Felidae, Canidae,
Equidae, Bovidae, Cervidae, Suidae, primates, rodents, Proboscidea,
Lagomorpha, Ursidae, pinnipeds, crustaceans, mollusks, cnidarians,
echinoderms, cetaceans, Mustelidae, Camelidae, Hippopotamidae]
Plants:
[flowers, trees, leaves, herbaceous plants, fungi, ferns, algae]
Food:
[dishes, desserts, beverages, fruit, ingredients, staple foods,
snacks, condiments]
Objects:
[tools, electronic devices, furniture, toys, containers, vehicles,
natural objects, musical instruments, sports equipment, stationery,
weapons, cultural artifacts, lighting fixtures, decorations]
Architecture:
[houses, towers, bridges, churches, temples, palaces, city walls, dams
or embankments, gate towers]
[Entities:  information and layout subjects]
Text:
[typographic lettering, sign lettering, title lettering, artistic
lettering, composited text, pages, articles, receipts, menus, resumes,
forms, manuals]
Posters:
[advertisements, flyers, covers, promotional materials, banners, film
or television works]
Interfaces:
[webpages, application interfaces, dashboards, screen interfaces,
panels]
Charts:
[bar charts, line charts, pie charts, tables, dashboards]
Signs and symbols:
[brand logos, organizational emblems, seals and stamps, flags and
heraldry, signage, interface icons, emoji, pictograms, glyph symbols]
[Scenes]
Indoor:
[bedroom, kitchen, office, hall, exhibition hall]
Landscape:
[mountains, forest, beach, river, desert, grassland]
```

Mapping prompt (continued)

```
Urban:
[street, skyline, city center, neon, square, park]
Space:
[galaxy, planet, nebula, Moon, astronomy]
Underwater:
[coral, reef, diving, seabed, ocean]
[Styles]
Painting:
[oil painting, watercolor, gouache, acrylic, ink painting, Chinese
painting, mural, printmaking, impasto, sketch, color work]
Cartoon:
[anime, comics, cute style, chibi, children's style, cel style, moe,
mecha, idol]
Rendering:
[computer graphics, models, assets, visualization, 3D]
Realism:
[photography, natural skin texture, realistic lighting and shadows,
everyday details, documentary feel]
Futurism:
[cyberpunk, steampunk]
```

3. Overall decision procedure

```
Make decisions in this order, but output only JSON:
1. Identify the image's main visual subject.
2. Determine whether that subject is an entity or primarily
represents a scene.
3. If there is a clear main subject, preferably output only the most
important entity.
4. Output two or three entities only when they are all clear,
independent, important, and comparable in prominence.
5. Do not enumerate background objects, supporting subjects,
decorations, or weakly visible elements.
6. If the image mainly conveys an overall environment, space,
landscape, or atmosphere without a single independent subject, the
entity list may be []; focus on the scene.
7. Scenes describe spatial environments, not every background
element.
8. Styles describe the overall visual medium or mode of expression,
not local details.
```

4. Entity hierarchy rules

```
Primary entity label:
Select strictly from the taxonomy's primary entity labels:
[Portraits, Characters, Activities, Clothing, Animals, Plants, Food,
Objects, Architecture, Text, Posters, Interfaces, Charts, Signs and
symbols]
Secondary entity label:
1. Prefer the secondary list under the corresponding primary label.
2. If all secondary candidates are clearly unsuitable, one stable,
general, reusable English secondary label may be added.
```

Mapping prompt (continued)

```
3. A new secondary label must be a category, not a specific instance,
brand, work, character, place, model, animal breed, or multiple
parallel terms.
4. Do not add a synonymous secondary label when an existing label
suffices.
Tertiary label:
1. Use a more specific subject name below the secondary label,
expressed naturally in English.
2. The tertiary label must not be identical to the primary or
secondary label.
3. Do not use specific animal breeds, model names, character origins,
work titles, or place names as tertiary labels; these belong in
factual labels.
4. Do not combine multiple subjects in one tertiary label.
5. Under portraits, publicly recognizable people belong to
celebrities rather than individual portraits. The tertiary label for
a celebrity should be the person's name.
6. For characters, the tertiary label defaults to the character or IP
name.
```

5. Factual label rules

```
A factual label must describe information that is directly supported
by the image, stable, reusable, finer than the tertiary label, and
discriminative of the subject's identity.
Permitted factual labels include:
1. Character origin: work title, series, game, brand, mythology, or
source culture.
2. Fine-grained animal identity: breed, species, subspecies, or
specific name, such as Labrador Retriever, Golden Retriever, Siamese
cat, or Bengal tiger.
3. A brand, organization, product series, or model with high public
recognizability and a stable product referent, clearly and directly
supported by the image.
4. If the tertiary label is already sufficiently clear, do not
redundantly define a factual label.
Leave the factual label empty when:
1. Evidence is insufficient.
2. It would be a guess based only on style, color, clothing, pose,
setting, or atmosphere.
3. It would repeat the tertiary label.
4. It is merely ordinary surface text, a slogan, number, shop name,
packaging text, sticker, painted graphic, low-recognition brand, or
low-recognition model.
5. It is merely a watermark, corner mark, small logo, background
text, or local marking.
6. The brand, character, animal breed, or specific location cannot be
established.
7. The subject is a portrait; portraits do not require factual
labels.
The factual label must be a single string, not multiple parallel
facts.
```

Mapping prompt (continued)

6. Subject priority and carrier rules

```
1. Prioritize the actually visible, physically independent main
subject.
2. Text, patterns, stickers, painted graphics, trademarks, and
character images on objects, clothing, packaging, screens, walls,
vehicles, or buildings are ancillary by default, not independent
subjects.
3. Output text, posters, screen content, charts, logos, or character
images as entities only when displaying that content itself is the
image's main purpose.
4. If a physical object is the main subject, do not separately output
character graphics, branding, text, or stickers on its surface.
5. Do not treat watermarks, corner marks, background logos, or
decorative text as entities.
6. Do not output both a carrier and its ancillary content unless both
are clear, independent, equally important subjects.
```

7. Category boundaries

```
[Portraits versus characters]
1. Portraits: the main subject is a real person.
2. Characters: the main subject is a stably recognizable character
or persona from film or television, or a fictional or non-real
character.
3. In real-world photographs, a real person defaults to portraits or
activities; do not directly assign characters.
4. For a real person in a character costume, cosplay, or
character-inspired styling, use portraits if the focus remains on the
real person.
5. Use characters only when the image clearly emphasizes the
character identity rather than the real individual.
6. Do not infer a specific character solely from clothing, hairstyle,
colors, props, pose, or atmosphere.
7. Put a specific character name in the tertiary label only with high
certainty.
8. If the originating work, series, game, or brand can be reliably
identified, put it in the factual label; otherwise leave it empty.
9. Do not guess specific names for generic cartoon figures, generic
anime figures, or weakly recognizable fictional figures.
10. Prefer the work or source for a character's factual label; put
the character name in the tertiary label.
[Secondary-label priority for characters]
1. Live-action film, television, or series characters: film or
television character.
2. Animation, comic, and derivative characters: animation or comic
character.
3. Characters clearly from a game context, when film/television or
animation/comics is less suitable: game character.
4. When both animation/comics and games are plausible, prefer
animation or comic character.
5. When both film/television and games are plausible, prefer film or
television character.
```

Mapping prompt (continued)

```
6. Character-based IP with stable public reach that does not fit
film/television, animation/comics, or games: popular IP.
7. Fictional, digital, or artistic personas that cannot be reliably
categorized: virtual persona.
[Animals]
1. An animal subject denotes the animal itself, not a character
identity.
2. Real animals, animal photographs, and ordinary animal
illustrations may all be animals.
3. Only stably recognizable fictional animal characters belong to
characters.
4. Prefer animal secondary labels from the taxonomy; add a stable
higher-level class if the existing secondary labels do not cover the
subject.
5. Common secondary-label examples:
  - Dogs, wolves, foxes: Canidae.
  - Cats, lions, tigers, leopards: Felidae.
  - Horses, zebras: Equidae.
  - Cattle, sheep, goats, antelopes: Bovidae.
  - Pigs, wild boars: Suidae.
  - Deer, Père David's deer: Cervidae.
  - Monkeys, apes: primates.
  - Mice, rats, squirrels, hamsters: rodents.
  - Elephants: Proboscidea.
  - Rabbits: Lagomorpha.
  - Bears, pandas: Ursidae.
  - Seals, sea lions: pinnipeds.
  - Whales, dolphins: cetaceans.
  - Otters, weasels, ferrets: Mustelidae.
  - Camels, alpacas: Camelidae.
6. Use familiar generic names as animal tertiary labels, such as dog,
cat, bird, horse, cattle, tiger, or dolphin.
7. Use factual labels for finer breeds, species, subspecies, or
specific names.
8. If the image clearly supports names such as Labrador Retriever,
Golden Retriever, Husky, Siamese cat, Ragdoll cat, Bengal tiger, or
Emperor penguin, put them in the factual label.
9. If only a generic identity such as dog, cat, or bird is visible,
the factual label must be "".
10. Do not put animal breeds in tertiary labels.
[Activities]
1. Activities emphasize a behavior, action, or event itself.
2. Use activities when people are playing sports, cooking, shopping,
performing, meeting, or gathering and the behavior is central to the
image.
3. For a static photograph of a person, prefer portraits.
4. Do not force an activity label merely because a few human actions
occur within an object or scene image.
[Clothing]
1. Clothing emphasizes garments, footwear, headwear, jewelry, or
outfit presentation itself.
```

Mapping prompt (continued)

```
2. Product images, outfit displays, and clothing-detail close-ups
preferably belong to clothing.
3. Prefer portraits when the real person is the focus and clothing is
merely what they wear.
4. Prints, text, and character graphics on clothing are not
independent entities by default.
[Plants]
1. Use plants when a single plant, flower, tree, leaf, fungus, or
similar subject is central.
2. For whole natural environments such as forests or grasslands,
prefer scenes; do not enumerate plants.
[Food]
1. Use food when meals, dishes, desserts, beverages, fruit, or
ingredients are central.
2. If food packaging is central, prefer objects or text; do not treat
packaging illustrations as actual food.
[Objects]
1. Use objects for general visible items.
2. Vehicles, electronic devices, furniture, tools, containers, toys,
and similar items belong to objects.
3. For dolls, figurines, and ornaments without a clear specific
character identity, prefer objects rather than characters.
4. Add an object's factual label only when its brand, model, or
series is very clear, highly recognizable, and sufficiently
recognizable to the public; otherwise leave it empty.
[Architecture versus scenes]
1. Use architecture when an individual building or landmark structure
is the main subject.
2. For an overall environment such as a city street, skyline,
building complex, or square, prefer the urban scene label Urban.
3. Put building names such as the Eiffel Tower, the Forbidden City,
or the Golden Gate Bridge in factual labels only with high certainty.
[Text / posters / interfaces / charts / signs and symbols]
1. Text: textual content or a readable layout is central, such as a
menu, form, receipt, or manual.
2. Posters: layout design is central, such as an advertisement,
cover, promotional material, banner, or film/television poster.
3. Interfaces: a webpage, application, dashboard, screen interface,
or control panel is central.
4. Charts: statistical or data presentation is central, such as a
bar chart, line chart, pie chart, or table.
5. Signs and symbols: a symbolic graphic is central, such as a brand
logo, organizational emblem, flag or heraldry, seal, sign, icon, or
emoji.
6. Do not output Logo or Icon as primary labels; use Signs and
symbols consistently.
7. For a complete poster, do not split its text, people, and icons
into multiple entities.
8. Prefer interfaces for a UI dashboard; prefer charts when a
statistical graphic or table is central.
9. Use signs and symbols when a brand logo is central; prefer the
product or vehicle when the logo is only a small mark on it.
```

Mapping prompt (continued)

8. Scene rules

```
Primary scene label:
Select from [Indoor, Landscape, Urban, Space, Underwater].

Secondary scene label:
1. Prefer the secondary list under the corresponding primary scene
label.
2. If all candidates are clearly unsuitable, a stable, general,
reusable English secondary scene label may be added.
3. The secondary label denotes a scene type, not a specific place,
city, country, or landmark. Put specific locations in factual labels.

Rules:
1. Scenes describe spatial environments, not individual objects.
2. Output the single most important scene by default.
3. For close-ups, object-only images, text-only images, icon-only
images, or product images on a plain white background without a clear
spatial environment, output [].
4. Scene factual labels contain only very clearly identifiable
places, landmarks, cities, countries, or scene identities.
5. If the specific location is uncertain, the factual label must be
"".
6. Do not infer locations from architectural style, clothing, written
language, or atmosphere.
```

9. Style rules

```
Primary style label:
Select from [Painting, Cartoon, Rendering, Realism, Futurism].

Secondary style label:
1. Prefer the secondary list under the corresponding primary style
label.
2. If all candidates are clearly unsuitable, a stable, general,
reusable English secondary style label may be added.
3. The secondary label denotes a medium, visual style, or mode of
expression, not a specific work, artist, brand, or subject matter.

Rules:
1. Style describes the overall visual medium or mode of expression.
2. Output the single most important style by default; [] is allowed
when evidence is weak.
3. Real-world photographs, documentary photographs, and images with
natural lighting and shadows: Realism / photography.
4. An obvious painting medium: Painting / the corresponding painting
type.
5. Generic cartoon, anime, comic, cute, or chibi appearance: Cartoon
/ the corresponding secondary label.
6. Obvious CG, 3D models, product renderings, or digital assets:
Rendering / the corresponding secondary label.
7. Obvious cyberpunk or steampunk visual language: Futurism / the
corresponding secondary label.
8. Text, charts, interfaces, and signs or symbols may have style []
if no clear medium or style is evident.
```

Mapping prompt (continued)

10. A few boundary examples

```
These examples illustrate the decision process, not the final output
format.  The final response must still contain only JSON.
Example 1:
The main subject is a clearly visible yellow Labrador Retriever.
Entity:  Animals / Canidae / dog / Labrador Retriever
Style, if it is a real photograph:  Realism / photography
Example 2:
The main subject in a product image is a T-shirt printed with an anime
character.
Entity:  Clothing / top / T-shirt / ""
Do not treat the print as a character entity.
Example 3:
Spider-Man is the main subject in an official poster or a character
still.
Entity:  Characters / film or television character / Spider-Man /
Marvel
Example 4:
The main subject is a standalone McDonald's Golden Arches logo.
Entity:  Signs and symbols / brand logos / McDonald's
Example 5:
A panoramic city-street view without one main building.
Entities:  []
Scene:  Urban / street / ""
```

11. Final output constraints

```
1.  Output valid JSON only.
2.  Do not output any explanatory text.
3.  Use English for all label values.
4.  Primary labels must come strictly from the taxonomy.
5.  Prefer taxonomy secondary labels; any added label must be a stable
category, not a specific instance or factual name.
6.  Output one entity by default, at most three; do not enumerate
every visible element.
7.  Output one scene by default; use [] if it cannot be determined.
8.  Output one style by default; use [] if it cannot be determined.
9.  Do not repeat identical, similar, or same-kind entities.
10.  Do not put secondary labels in primary-label fields.
11.  An entity's tertiary label must not be identical to its primary
or secondary label.
12.  Do not put the main-subject name that belongs in the tertiary
label into the factual label, especially character names, celebrity
names, dish names, building names, or logo names.
13.  Do not use factual labels as tertiary labels, especially animal
breeds, character origins, or work titles.
14.  When the factual label is uncertain, fill in "".
15.  Do not include the words unknown, other, none, uncertain, or
possibly in the output.
16.  The final response must start with { and end with }; do not use a
code block.
```

C HIERARCHY ANALYSIS

The following diagnostics assess semantic organization, clustering sensitivity, and category memberships at complementary levels.

C.1 SEMANTIC ORGANIZATION OF THE RETAINED HIERARCHY

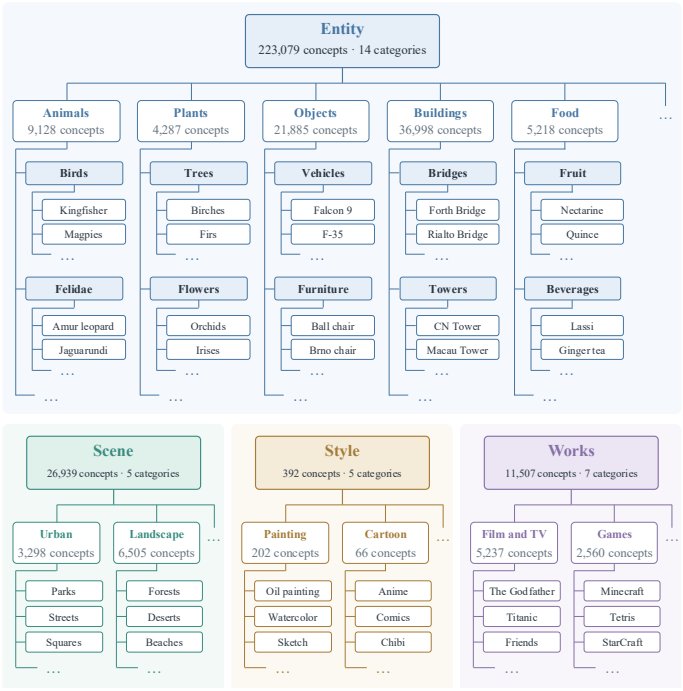


Figure 6: Example hierarchy paths; subtree counts can overlap across branches.

Measured results. We evaluate 246 L2 categories in 23 L1 branches using 1,024-dimensional concept embeddings (Table 7). All 23 branches exceed the permutation baseline in mean silhouette; the L2 macro-averages are 0.097 and -0.139, respectively (Figure 7a). At $k = 5$, same-category neighbor precision is 55.5%, compared with 9.0% under permutation (Figure 7b). The 46.5 percentage-point difference supports local semantic organization beyond what category sizes alone explain. Triplet concordance is 76.5% at L2, 68.5% at L1, and 49.2% at the aspect level against a 50% reference (Figure 7c).

Table 7: Measured graph diagnostics against permutation and chance references.

	Silhouette	Neighbor	Triplet concordance (%)		
Memberships	(cosine)	P@5 (%)	L2	L1	Aspect
Hierarchy	0.097	55.5	76.5	68.5	49.2
Reference	-0.139	9.0	50.0	50.0	50.0

Membership and sampling. We use original category memberships for concepts retained in the final graph, merging aliases by canonical node ID. Of 470 native L2 candidates, 280 contain at least five members with available embeddings, totaling 54,027 members. These categories form the L2 affinity matrix in Figure 8a. Metrics conditioned on a parent category require a unique L1 parent and at least one eligible sibling category, yielding 246 L2 categories in 23 L1 branches. The affinity matrices use all eligible members. Silhouette and neighbor evaluations draw up to 40 members per

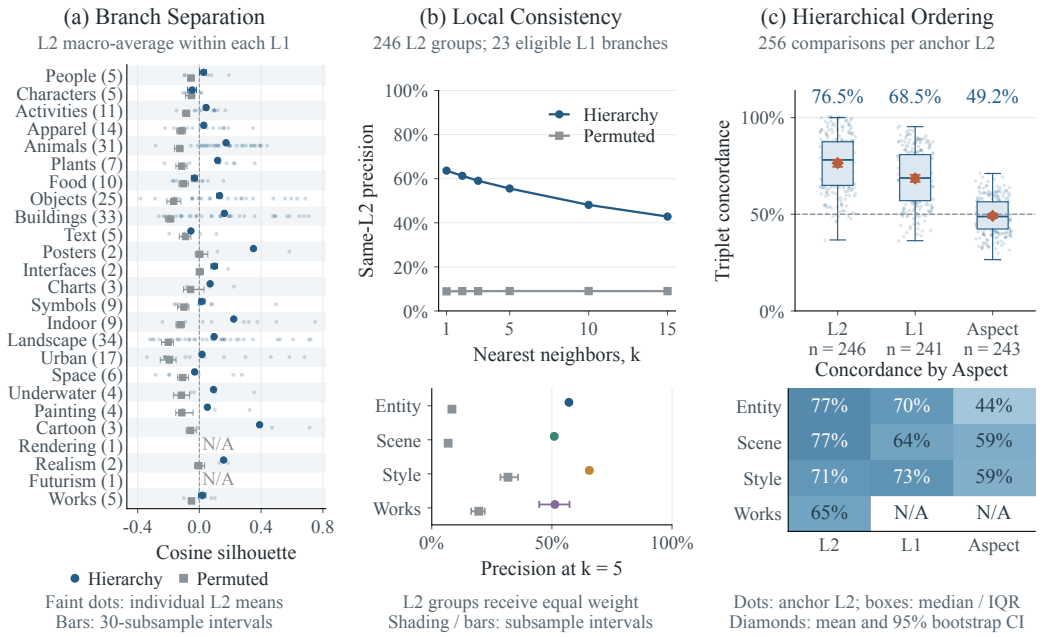


Figure 7: Measured semantic organization of the retained hierarchy. (a) Cosine silhouette by L1 branch; faint dots show L2 means. (b) Same-L2 neighbor precision across neighborhood sizes and by aspect at $k = 5$. (c) Triplet concordance at L2, L1, and aspect levels. Intervals in (a,b) span the 2.5th–97.5th percentiles of 30 subsamples; those in (c) are 95% bootstrap intervals. N/A denotes an inapplicable comparison or insufficient eligible groups.

L2 category in each of 30 repetitions, with sampling seed 50. Assignments remain fixed throughout; these repetitions quantify sampling variation rather than sensitivity to clustering initialization.

Metric definitions. Silhouette uses cosine distance to compare cohesion within an L2 category with separation from its siblings. Neighbor precision is the fraction of the k nearest other concepts in the same L1 branch that share the anchor L2 label. Both scores are averaged within each L2 category and then equally across L2 categories. The permutation baseline preserves category sizes within each L1 branch. Intervals for these two metrics are the 2.5th–97.5th percentiles across subsamples.

L2 triplets compare a same-L2 positive with a different-L2 negative under the same L1 parent. L1 triplets use a different-L2 positive under the same parent and a different-L1 negative in the same aspect. Aspect triplets use a different-L1 positive in the same aspect and a negative from another aspect. Each eligible anchor category contributes 256 triplets; identical anchor–positive concepts are excluded, and ties receive 0.5. The three comparisons have 246, 241, and 243 eligible anchor categories, with mean concordances of 76.5%, 68.5%, and 49.2%, respectively. The 95% intervals use 2,000 bootstrap resamples of anchor categories. The L2 and L1 intervals are [74.6, 78.2]% and [66.7, 70.4]%, respectively.

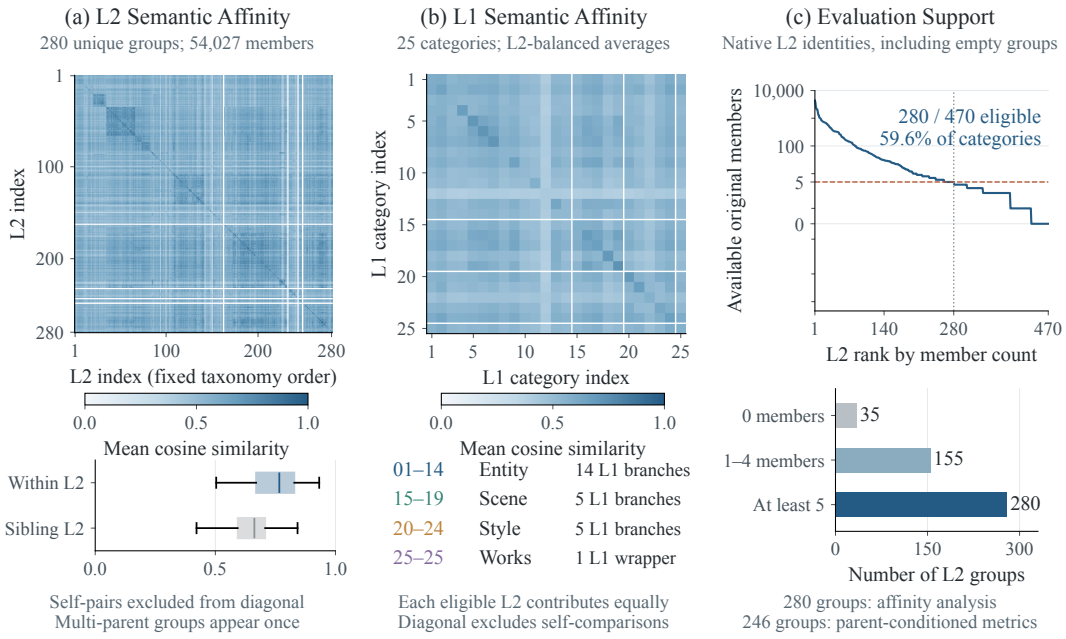


Figure 8: Measured category affinities and evaluation coverage. (a) Mean cosine affinities of 280 L2 categories in taxonomy order, with within-category and sibling-category distributions. (b) L1 affinities, weighting eligible L2 categories equally within each parent. Matrix diagonals exclude self-pairs. (c) Member support across 470 L2 candidates: 280 qualify for the affinity analysis, and 246 qualify for parent-conditioned metrics.

C.2 Clustering Stability and Granularity

Inputs and hierarchy. We extract 1,024-dimensional embeddings for 706 concepts using jina-embeddings-v5-text-small (Akram et al., 2026). The set contains 277 entities, 181 scenes, 152 styles, and 96 works. All input embeddings are ℓ_2-normalized before clustering. We construct a separate hierarchy for each aspect using bottom-up hierarchical k-means (Vo et al., 2024). We first cluster concept embeddings at the finest level, then recursively cluster the resulting centroids to form higher levels. Assignments at each level are propagated to the original concepts. Upper stages weight their input centroids equally without renormalization. Each stage uses Euclidean k-means with k-means++ initialization and ten initializations per fit. The iteration limit is 300, and the convergence tolerance is 10^{-4}.

Seed stability and granularity. We assess initialization sensitivity using seeds 0–49. Input embeddings and cluster counts remain fixed, and each run rebuilds the full hierarchy. For each run, we compare its L2 assignments with those from the other 49 runs using the adjusted Rand index (ARI). Figure 4(b) reports the mean ARI for each run in seed order. These scores share comparison runs and describe initialization sensitivity rather than confidence intervals.

The granularity analysis varies the finest-level cluster count from approximately 0.5 to 2 times its reference value. Counts are restricted to integers and must exceed the next coarser count. Each setting uses seeds 0–9 while keeping upper-level counts fixed. Silhouette scores use cosine distance and are averaged over the original concept embeddings. The singleton fraction is the proportion of clusters containing one concept. Bands in Figure 4(a) show the 10th–90th percentiles across seeds, not confidence intervals.

Membership consistency. We use seed 0 as the fixed reference for cluster-level comparisons. For each reference cluster, we compute its highest Jaccard similarity to a cluster at the same level in each of the other 49 runs. Jaccard similarity compares the sets of concept members. Single-cluster levels and undefined comparisons are reported as N/A.

C.3 Human Review of Category Memberships

To stress-test the learned category memberships, we focus on boundary concepts rather than prototypical cluster members. For each of the 246 eligible L2 categories, we rank concepts by the margin between their similarity to the assigned cluster centroid and their highest similarity to a sibling-cluster centroid, and select the three concepts with the smallest margins, yielding 738 concepts. Two reviewers independently assign each concept to a candidate L2 category without seeing its original graph membership, with uncertain cases recorded separately. Figure 9 compares cosine silhouette scores under the original graph memberships and the human-reviewed assignments across 23 eligible L1 branches. The similar branch-level patterns indicate that the learned hierarchy remains semantically consistent with human judgments even for ambiguous concepts near category boundaries. Rendering and Futurism each contain only one eligible L2 category, with 12 and 16 members, respectively, and are excluded because within-parent silhouette requires at least two eligible categories.

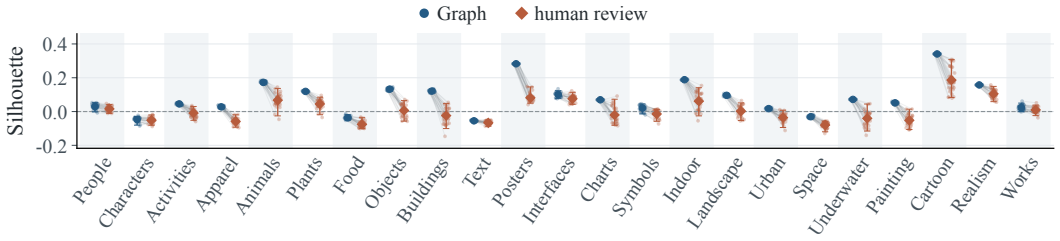


Figure 9: Cosine silhouette of graph and human-reviewed memberships across 23 L1 branches. Blue circles: graph; orange diamonds: human review.

D KnowGapBench Construction

D.1 Task Composition and Concept Distributions

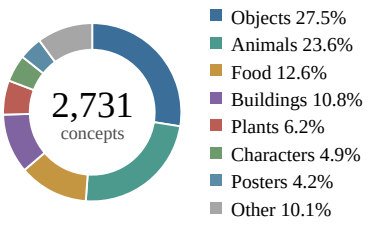
Table 8 summarizes the KnowGapBench task settings from Section 4.2.

Table 8: KnowGapBench task composition. Entity-task budgets exclude the global scene or style condition.

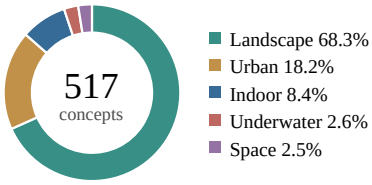
Task	Targets	Budget	Prompts
CR	1	1	973
CI	1	2	976
SC	1	3	1,149
JC	2	3	600
Total	—	—	**3,698**

Figure 10 presents the four branches of the benchmark concept inventory. Taxonomy and membership edges provide named subdivisions for 3,256 of the 3,286 targets, covering 30 hierarchical categories: 13 entity, five scene, five style, and seven work categories. The category plots show these observed subdivisions; 30 targets without named subdivisions remain in the benchmark but are excluded from the plots. These hierarchical categories describe the inventory, while the ten semantic families in Table 12 define the evaluation aggregation.

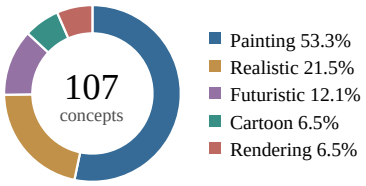
The named subdivisions contain 2,731 entity, 517 scene, 107 style, and 370 work memberships. Memberships can overlap across branches. Within each branch, every canonical concept contributes unit weight, divided equally among its recorded categories when it has multiple memberships. Percentages use the corresponding branch total and are rounded for display.



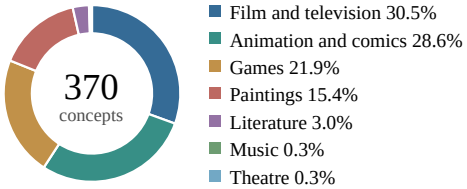
(a) Entities: 2,731 memberships.



(b) Scenes: 517 memberships.



(c) Styles: 107 memberships.



(d) Works: 370 memberships.

Figure 10: KnowGapBench concept distributions across the entity, scene, style, and work hierarchies. Percentages are normalized within each branch; concepts may belong to multiple branches.

D.2 Reference Images

Figure 11 shows four reference images from the released benchmark library, selected for clear subjects and visible detail. They provide supporting evidence for Knowledge evaluation after reference auditing.



(a) Airedale Terrier.



(b) Green Bank Observatory.



(c) Cape Eluanbi.



(d) Tessellation.

Figure 11: Examples of benchmark reference images. Panels illustrate an animal identity, an observatory structure, a landmark scene, and a geometric pattern. References support identity judgments without requiring generated images to reproduce their viewpoint or composition.

E BENCHMARK EVALUATION

E.1 EVALUATION PROCEDURE

We evaluate generated images with Qwen3.8-27B (Qwen Team, 2026a) using the frozen English evaluator prompts (`en_v3`). Table 9 summarizes the evidence and scoring roles of each evaluator call.

Reference audit. Before image scoring, a separate evaluator checks each selected reference against the target's bilingual names, identifier, role, and graph context. A reference can reach the Knowledge evaluator only when its audit status is `matched` or `partial`, confidence is at least 0.70, and the target is the main subject; benchmark preparation also checks Parent compatibility for entity targets. The reference-audit prompts are inherited from `en_v2`.

Independent judgments. Each requirement is evaluated in a separate conversation, without other scores or the task-complexity label. Parent checks the broad category, Knowledge checks target identity with audited reference evidence, and Object Presence, Attribute Binding, and Relation assess composition using the generated image as their visual input. For attributes and relations, distinct participants are matched using category and independent visible cues, without using the property being tested to choose the match.

Table 9: Evaluator inputs and scoring roles. All image judgments receive the generated image; only Knowledge additionally receives eligible reference pixels.

Judgment	Additional evidence or requirement	Raw scale / role
Parent	Target name and broad category	0–4; entity prerequisite
Knowledge	Identity, graph context, audited reference	0–4; concept identity
Object Presence	Object question, criterion, graph context	0 or 4; auxiliary presence
Attribute Binding	Attribute question and criterion	0 or 4; composition
Relation	Directed relation question and criterion	0 or 4; composition
Style	Global scene, medium, lighting, or style condition	0–4; condition compliance
Visual quality	Visible defects and overall usability	0–4; diagnostic only

Normalization and dependencies. Raw scores are divided by four, and prerequisites are applied using Equation 5. Entity Knowledge depends on Parent; attributes and relations depend on participant presence (Parent for targets and Object Presence for auxiliaries). Fine-grained identity errors alone do not invalidate correct composition. The evaluator can return `not_judgeable` when visibility or ambiguous instance assignments prevent a sound decision; unavailable scores are excluded while observed and gated zeros are retained.

E.2 VISUAL RELATIONS

Table 10 lists the nine spatial relations selected in the frozen benchmark, totaling 466 relation requirements. Each evaluator question preserves the specified participants and direction; these image-level requirements are scored separately from graph facts used for identity disambiguation.

Table 10: Spatial relations used in benchmark composition and evaluation. Counts refer to relation requirements, pooled across SC and JC.

Family	Relation wording	Count
Horizontal	located to the left of	47
Horizontal	located to the right of	50
Vertical	located above	57
Vertical	positioned below	38
Depth	positioned in front of	51
Depth	located behind	60
Adjacency	located next to	46
Proximity	located near	64
Proximity	distant from	53
Total		466

E.3 Evaluator Prompts

The boxes below reproduce the English prompts used by the evaluators; dollar-prefixed fields are filled from the target record or the corresponding requirement. The reference audit has its own system prompt, while the image judgments share one system prompt.

E.3.1 Reference-Image Evaluator

Reference evaluator: system prompt

```
You are the final reference-image auditor for a benchmark. Verify
whether one image already selected from a retrieval pool can serve as
supporting evidence for the Knowledge score. Do not rerank the pool
and do not evaluate a model-generated image. Independently identify
the target using both its English and Chinese names, exact knowledge
identifier, supplied knowledge-graph context, source visual-type hint,
and reliable world knowledge. The source visual type is coarse
project metadata and may be corrected by stronger identity evidence,
but it is independent of the candidate pixels. Never reinterpret a
transliterated English label as a different homonym when the Chinese
name or identifier resolves the target. A web image may show only
part of the target, place it away from the main subject, or depict the
wrong concept. Never assume it is correct merely because it was
supplied. Return JSON only.
```

Reference evaluator: user prompt

```
Target English name: $concept_name_en
Target Chinese name: $concept_name_zh
Knowledge identifier: $concept_id
Evaluation role: $benchmark_role
Reliable aliases: $aliases
Confusable concepts (not aliases): $confusable_with
Definition or page summary: $definition
Knowledge-graph context: $concept_context
Source visual-type hint (Chinese): $source_visual_type_hint
Fine category: $fine_category
Expected broad visual category: $parent_category
Taxonomy policy: $taxonomy_policy_version
```

Reference evaluator: user prompt (continued)

```
Decide how the candidate image relates to the target.  For an entity,
the image should show the concept itself.  For a work, the image may
show a representative character, scene, or prop supported by an
explicit one-hop fact.  For an activity, it should show a recognizable
combination of participants, actions, props, and setting.  For a
scene, the target environment, spatial layout, and characteristic
components should dominate.  For a visual style, the rendering
characteristics should govern the whole image rather than appear only
as written text or a small local patch.  For a work,
target_is_main_subject asks whether the supported representative
character, scene, or prop is the main subject; it does not require an
abstract work to become an object.  For a scene or style, it asks
whether that environment or rendering style is the dominant evidence.
If target evidence is reliably identifiable but only local or not the
main subject, use partial and target_is_main_subject=false.  Use
mismatch for a clear error and ambiguous when verification is not
possible.

Only for the entity role, also verify whether the expected broad
visual category is a compatible closed-set operational Parent.
Determine compatibility from the bilingual target identity,
identifier, KG, source visual type, and reliable knowledge--not from
what the candidate image happens to depict.  A valid fine category
must remain compatible with its broader parent; for example, an
instrument is compatible with tools and equipment, an island or cape
with natural objects, and a dish with food or drink.  Do not mark a
parent incompatible merely because a more specific category exists.
Mark it incompatible only when the target identity genuinely
contradicts the broad category.  For work, activity, scene, and style
roles, parent_status must be not_applicable.

Output:  {"status":"matched|partial|mismatch|ambiguous","confidence":
0.0,"target_is_main_subject":true,"parent_status":"compatible|
incompatible|uncertain|not_applicable","suggested_parent":"empty when
compatible or not applicable","usable_visual_evidence":["up to four
stable visible features"],"reason":"50 words or fewer"}
```

E.3.2 Identity and Composition Evaluators

Image evaluator: shared system prompt

```
You are the visual judge for a text-to-image benchmark.  Evaluate only
the single dimension requested in the current message.  Do not infer
or adjust scores for any other dimension or for criteria that are not
shown.  Base the decision strictly on visible evidence and return only
the requested JSON object, with no explanatory prose.
```

Image evaluator: Parent

```
Does the evaluated image clearly show the subject required by
"$concept_name", and does that subject belong to the broad visual
category "$parent_category"?

Judge only subject presence and the coarse category.  Do not require
fully correct fine-grained identity, and do not judge composition,
style, or image quality.

Scores:  0=missing subject or wrong broad category; 1=only weak cues;
2=ambiguous subject or category; 3=mostly clear; 4=clear and correct.
Output:  {"score":0,"confidence":0.0,"status":"scored|not_judgeable",
"evidence":"30 words or fewer of visible evidence"}
```

Image evaluator: Knowledge

```
Does the evaluated image recognizably depict the concept
"$concept_name" (knowledge identifier:  $concept_id; role:
$benchmark_role)?

Concept text:  $definition
Reliable aliases:  $aliases
Wikidata one-hop context (fact support only; it need not all be
depicted):  $wikidata_context
Confusable concepts:  $confusable_with
Candidate Concept Card visual cues (must be verified):
$visible_anchors
Candidate reference-image audit:  $reference_audit
Category-specific observation guidance:  $category_guidance

Rules:
1.  A web reference image is revocable supporting evidence, not
absolute ground truth, and the evaluated image need not copy its
background, pose, viewpoint, lighting, or composition.  Prioritize
reliable world knowledge, the concept name, its identifier, and
compatible Wikidata relations.  Ignore conflicting evidence.
2.  Judge an entity by stable visible identity features.  A work is
not established merely by showing its title or logo; use reliably
related characters, scenes, props, or art direction.  Judge an
activity from the combination of participants, actions, typical props,
and setting, not from a text sign alone.  Judge a scene by its overall
environment, layout, and characteristic components.  Judge a visual
style from consistent image-wide rendering features, not from a
written style name.
3.  Wikidata context is for disambiguation.  Invisible facts such as
abstract type, creator, origin, or date cannot by themselves prove
that an image is correct.  Only graph relations consistent with
visible content count as positive evidence.
4.  Use not_judgeable when neither the image nor reliable knowledge
permits a sound decision.  Do not guess.

Scores:  0=clearly not the concept; 1=only broadly similar subject
matter; 2=some key cues are correct but identity remains ambiguous;
3=recognizable with minor omissions; 4=clear concept identity with
consistent key evidence.
Output:  {"score":0,"confidence":0.0,"status":"scored|not_judgeable",
"reference_use":"used|partial|ignored|none","evidence":"40 words or
fewer of visible evidence"}
```

Image evaluator: Object Presence

```
Judge only whether the following object is visibly present and has the
correct identity in the evaluated image:
$question
Criterion: $criterion
Wikidata or taxonomy context for disambiguation: $wikidata_context
The context may disambiguate a name but cannot replace visible
evidence, and invisible knowledge need not be depicted. Give 4 when
the requirement holds and 0 when it does not. Use not_judgeable only
when occlusion or the image genuinely prevents a decision. Ignore all
other scoring dimensions.
Output: {"score":0,"confidence":0.0,"status":"scored|not_judgeable",
"evidence":"30 words or fewer of visible evidence"}
```

Image evaluator: Attribute Binding and Relation

```
Judge this single visible requirement:
$question
$criterion
Match distinct object instances using visible category and independent
cues, never the attribute or relation being tested. Fine-grained
identity mismatch alone is not a failure.
Give 4 when the requirement holds; give 0 when a participant is
missing, its visible category is wrong, or the requirement fails. Use
not_judgeable only when visibility prevents a decision or plausible
instance assignments yield different outcomes.
Output: {"score":0,"confidence":0.0,"status":"scored|not_judgeable",
"evidence":"30 words or fewer of visible evidence"}
```

For Attribute Binding, the criterion is "The specified attribute must hold on the corresponding object instance." For Relation, it is "The specified relation must hold between the corresponding object instances, with the stated direction." These are the criterion substitutions used by `en_v3`.

E.3.3 Global Condition and Diagnostic Quality Evaluators

Image evaluator: global condition (Style)

```
Judge only whether the evaluated image satisfies this scene, medium,
lighting, or visual-style requirement:
$style_requirement
Do not judge the target's knowledge identity, object relations, or
general image quality.
Scores: 0=clear conflict or absent; 1=weakly related; 2=partially
satisfied; 3=mostly satisfied; 4=clearly satisfied.
Output: {"score":0,"confidence":0.0,"status":"scored|not_judgeable",
"evidence":"30 words or fewer of visible evidence"}
```

Image evaluator: diagnostic visual quality

```
Evaluate only the general visual quality of the image itself,
including severe artifacts, broken structures, unreadable gibberish,
abnormal cropping, and overall usability.  Do not judge whether it
follows the prompt or depicts a particular knowledge concept.

Scores:  0=unusable; 1=severe defects; 2=clearly defective but
recognizable; 3=good quality with minor flaws; 4=clear, complete, and
free of obvious defects.

Output:  {"score":0,"confidence":0.0,"status":"scored|not_judgeable",
"evidence":"30 words or fewer of visible evidence"}
```

The visual-quality call is diagnostic; the reported preference metric is the independently computed HPSv3 (Ma et al., 2025).

E.4 Evaluator Validation

Table 11 reports agreement between the automated evaluator and human judgments for concept identity and compositional requirements.

Table 11: Evaluator comparison on 200 samples. Best in bold; second best underlined.

	Knowledge		Atomic judgments	
Evaluator	$\rho\uparrow$	$\kappa_w\uparrow$	Acc. $\uparrow$	$\kappa\uparrow$
Qwen3.8-27B	0.704	<u>0.772</u>	80.0	<u>0.581</u>
Qwen2.5-VL-72B	0.610	0.687	78.1	0.546
GPT-5.6-Sol	**0.818**	**0.813**	**83.3**	**0.610**
Gemini 3.1 Pro	<u>0.786</u>	0.769	79.7	0.579
GLM-5.3-Flash	0.692	0.685	<u>80.4</u>	0.575

E.5 Aggregation and Detailed Results

The tables use the completed evaluations from the frozen Rare release with Qwen3.8-27B. Each prompt has one generated image when generation succeeds. The reported means exclude unavailable scores but retain observed zeros, including zeros introduced by dependency gating.

Parent applies to entity targets; Knowledge also includes other semantic roles. Their overall means therefore need not use the same samples. The matched comparison in Section 5.3 restricts both metrics to entity prompts with both scores available. In JC, the two Knowledge scores are averaged, whereas Parent is the minimum of the available target Parent scores. JC does not measure simultaneous success on both target identities. Overall in Table 2 is the Knowledge concept-macro score computed from all available CR, CI, SC, and JC records before rounding. Each individual concept is one averaging unit and each JC concept pair is one unit. Overall is therefore not the unweighted mean of the four displayed task scores and does not combine Parent, Knowledge, and Atomic scores.

The atomic components in Table 3 are recomputed from the pooled SC and JC samples, not averaged from rounded task scores. Style is the existing global condition compliance score; HPSv3 (Ma et al., 2025) is an independent visual preference score on its native scale. Table 12 uses the ten evaluation semantic categories for the 2,086 individual targets in CR, CI, and SC. The 600 JC pairs, comprising a further 1,200 targets, are reported separately in Table 2. These evaluation categories should not be interpreted as the four top level branches of the concept hierarchy.

Model names are shortened in the tables. Gemini 3.1 Flash denotes Gemini-3.1-Flash-Image-Preview; Seedream 5.0 denotes Doubao-Seedream-5.0-260128; FLUX.2 Pro denotes Flux-2-Pro; HunyuanImage-3.0 denotes HunyuanImage-3.0-Instruct; and SenseNova-U1.5 denotes SenseNova-U1.5-8B-MoT. UniWorld Fine-T2I denotes the Fine-T2I dataset baseline. All other names follow the evaluation records.

Table 12: Knowledge scores by semantic category, aggregated over CR, CI, and SC. JC is excluded because its evaluation unit is a concept pair rather than an individual category. Higher is better.

Model	Bio.	People	Vehicles	Built	Everyday	Symbols	Works	Activities	Scenes	Styles
GPT-Image-2	39.3	**49.1**	**70.5**	**54.6**	**54.4**	**62.6**	**60.5**	**73.1**	**63.6**	**63.0**
Gemini 3.1 Flash	**44.9**	31.4	62.4	50.7	52.1	49.2	47.9	71.1	55.1	62.3
Seedream 5.0	29.5	35.4	47.0	40.4	44.9	44.4	37.5	59.2	42.5	54.4
FLUX.2 Pro	30.0	21.9	49.2	33.7	43.4	32.1	21.5	56.8	38.6	57.0
Qwen-Image-2.0-Pro	30.8	27.7	41.4	32.9	40.3	39.6	27.1	65.4	34.3	49.3
HunyuanImage-3.0	20.9	28.0	41.0	27.8	32.2	31.1	27.9	49.6	25.1	41.8
Boogu-Image	15.2	19.6	31.2	23.1	26.3	20.9	20.3	48.7	33.5	49.5
SenseNova-U1.5	14.1	14.1	31.4	23.7	23.8	23.0	19.6	49.6	26.6	50.4
Z-Image	11.5	18.4	28.4	20.9	23.6	26.2	23.8	40.4	26.7	46.7
Qwen-Image	14.8	15.8	28.5	21.6	23.8	26.4	23.1	38.6	21.6	38.6
UniWorld Fine-T2I	8.8	12.0	18.9	18.0	16.8	15.5	14.0	35.1	19.5	31.7
GLM-Image	8.9	15.3	21.7	13.8	18.8	12.6	16.3	36.0	19.2	21.6
InternVL-U	3.9	10.6	12.7	10.3	10.0	10.0	10.8	21.1	14.4	31.9

F CGTR TRAINING AND EVALUATION

F.1 TRAINING SETTINGS AND SAMPLING PRIORS

Table 13 lists the UniWorld-V1 fine-tuning settings for the Fine-T2I baseline and CGTR refinement. Both runs process 6,177,306 training entries in one epoch. Their global batch sizes differ, so equal sample exposure entails different numbers of optimizer updates.

Table 13: Fine-tuning settings for UniWorld-V1.

Setting	Fine-T2I baseline	CGTR refinement
Initialization	UniWorld-V1	UniWorld-V1
Fine-T2I entries	6,177,306	5,841,277
Supplementary entries	0	336,029
Total entries	6,177,306	6,177,306
Supplementary share	0%	$\approx 5\%$
Epochs	1	1
Optimizer updates	96,521	24,131
Global batch size	64	256
GPUs / batch per GPU	16 / 4	64 / 4
Gradient accumulation	1	1
Warmup updates	4,826	1,206
Learning rate	5×10^{-6}; cosine decay	
Optimizer	AdamW; $\beta = (0.9, 0.99)$, $\epsilon = 10^{-8}$, weight decay 0	
Training resolution / precision	512×512 / bfloat16	
Training seed	203	

Dataset composition. The training mixture produced by CGTR contains 5,841,277 resampled Fine-T2I entries and 336,029 supplementary image–caption pairs, with approximate shares of 95% and 5%. The supplementary collection comprises 187,714 pairs from the initial gap-concept collection and 148,315 from newly identified gaps. Each supplementary pair enters the mixture once, without a separate sampling prior for either source.

Category priors. Table 14 gives the predefined priors for resampling the Fine-T2I portion of the mixture, normalized to 100%. Each record receives its first recognized primary entity label; records without one enter the Objects group, making the groups disjoint. We multiply each prior by 5,841,277, round down, and allocate the remaining entries by descending fractional remainder. Groups are sampled without replacement when they meet their quota and with replacement otherwise. Thus, training-entry counts can include repeated images. The selection uses seed 20260918 and is shuffled during training; the distributed loader may repeat entries to fill the final batch.

F.2 COMMONCONCEPTBENCH EVALUATION

CommonConceptBench contains 1,000 bilingual prompts selected using observed concept frequency and manual review: 300 CR, 350 CI, and 350 SC prompts. CI and SC can contain multiple familiar targets, with no separate JC setting. Prompts follow the graph-based construction procedure of KnowGapBench, including consistency checks, a caption reasonableness threshold of 0.70, and translation review. All targets have audited reference images and one-hop graph context.

The 632 concepts include secondary targets in multi-concept prompts. Grouping prompts by primary concept yields 463 averaging units overall and 300, 326, and 275 within CR, CI, and SC. For multi-target prompts, Knowledge is the minimum target score before concept-macro aggregation. KnowGapBench JC instead averages two target scores, so absolute scores across the two benchmarks do not isolate the effect of concept familiarity. Overall pools Knowledge scores before rounding and is not the mean of the displayed task scores. Both models are evaluated with Qwen3.8-27B (Qwen Team, 2026a) and the en_v3 judging prompts using identical published reference-audit decisions.

Table 14: Category sampling priors for CGTR.

Primary category	Prior (%)	Training entries
People	16	934,604
Objects	14	817,779
Text	13	759,366
Clothing	12	700,953
Activities	10	584,128
Architecture	7	408,889
Food	5	292,064
Characters	5	292,064
Animals	5	292,064
Plants	4	233,651
Signs and symbols	3	175,238
Posters	3	175,238
Charts	2	116,826
Interfaces	1	58,413
Total	100	5,841,277

An overlap audit finds no shared target identifiers between CommonConceptBench and the supplementary training annotations, or exact caption matches after Unicode normalization, case folding, and whitespace normalization. Both English and Chinese benchmark prompts are checked against the supplementary captions. Common concepts may still appear in Fine-T2I or as auxiliary content in supplementary images. The benchmark therefore tests transfer beyond the targets selected for supplementation, rather than concepts unseen throughout training.