EduT2I: A Content-Stratified Benchmark for Caption-to-Diagram Generation in Secondary-School STEM

Anonymous ACL submission

Abstract

We introduce EduT2I, a caption-to-diagram benchmark drawn from publicly available Indian secondary-school textbooks across physics, chemistry, biology, mathematics, and general science. The corpus contains 1,685 figures across 158 chapters, with caption coverage of 82.6% after a two-pass extraction step. We tag every chapter with one of 35 frozen content-area labels using a deterministic keyword-rule pass and a local-LLM fallback, then stratify the train/val/test split on the tag rather than on the (class, subject) key. We use the benchmark to evaluate three publicly available text-to-image generators (Stable Diffusion 1.5, SDXL-Turbo, SDXL-base at 512) in a zero-shot setting and two LoRA-fine-tuned variants of Stable Diffusion 1.5 trained on AI2D and on AI2D+ScienceQA. All five models underperform on standard image-quality metrics by a wide margin (best aggregate FID 133.7 at n=478), and the two fine-tuned variants *regress* every metric versus the zero-shot model they were built on top of. The regression is larger when the fine-tuning corpus mixes AI2D with ScienceQA than when it uses AI2D alone, and bootstrap 95% confidence intervals on the per-sample metrics confirm the gaps are not noise. We release the corpus, the tagging pipeline, the split manifest, the evaluation harness with bootstrap CIs, and 2,390 generated panels so future work can target the specific failure modes the benchmark documents.

1 Introduction

Educational diagrams sit in an awkward gap inside the modern text-to-image literature. They are not natural photographs, where most contemporary image generators have been tuned, and they are not synthetic icons, where rasterisation pipelines such as TikZ already do a perfect job when supplied with the right code. They are stylised illustrations: a textbook figure of cellular respiration has a fixed colour grammar, a level of abstraction chosen for a ten-year-old, and an annotation density that natural-image generators have no prior on. Producing such figures from a free-text caption is closer to writing structured graphics than to rendering a scene. It is also a problem worth posing because most of the world's school content is currently locked into static PDFs, and the working hypothesis behind every text-to-image-for-education project is that those PDFs can be regenerated, localised, and edited as easily as a paragraph of Wikipedia.

The benchmark literature has not caught up. AI2D (Kembhavi et al., 2016) and its extensions stop at ≈5,000 elementary-school diagrams, all in English and all reusing a small visual vocabulary. ScienceQA (Lu et al., 2022) is broader in question types but the underlying diagram corpus is again English-only and dominated by mid-school US content. No benchmark we are aware of targets late-secondary STEM diagrams from a non-Western curriculum, includes both the source captions and the figures at chapter granularity, and ships a frozen split that can be used for zero-shot evaluation of a modern text-to-image stack.

EduT2I fills this gap with material from the National Council of Educational Research and Training (NCERT) textbook series used in Indian classes 6 through 12. The books are public, licensed for educational reuse, cover a curriculum that takes physics, chemistry, biology, and mathematics through the AP-equivalent level, and present content with the dense, labelled diagram style we want to study. The dataset we release contains 1,685 figures, 1,392 free-text captions, 158 chapters, and a two-pass content-area tagging system that lets us stratify the train/val/test split on the actual subject matter of each chapter rather than on the textbook it came from.

We use the benchmark to make three measure-

ments. *First*, off-the-shelf text-to-image performance on Indian secondary-school content: we evaluate Stable Diffusion 1.5, SDXL-Turbo, and SDXL-base at 512 in a zero-shot setting on a stratified sample of 478 figures spanning all five subjects. *Second*, whether neighbour diagram corpora help: one LoRA is trained on AI2D + ScienceQA, one on AI2D alone, both on top of the same SD 1.5 base. *Third*, we report 95% bootstrap confidence intervals at 1,000 resamples on every per-sample metric so that headline claims have uncertainty bars attached.

The result is uncomfortable but instructive, and we frame the paper explicitly as a negative-result contribution. Every model we test sits in FID range 134–202, an order of magnitude above where SD 1.5 itself was tuned on natural images. Fine-tuning on neighbouring diagram data does not narrow this gap: the combined-corpus LoRA is the *worst* performer on FID, KID, and CLIPScore; the AI2D-only LoRA is second-worst. The distilled SDXL-Turbo is the best of the five, beating its own full-sized parent. We read this as evidence that the failure mode is upstream of fine-tuning: the model's prior over what an "educational diagram" looks like does not match the NCERT visual idiom, and pulling the prior further toward AI2D (a different idiom again) makes the mismatch worse, not better. The benchmark exists in part to make this kind of negative finding quantitative and reproducible.

2 Related work

Diagram corpora. The closest neighbours of EduT2I are AI2D (Kembhavi et al., 2016), AI2D-RST (Hiippala et al., 2021), and the diagram subsets of ScienceQA (Lu et al., 2022) and TQA (Kembhavi et al., 2017). AI2D contains 5,000 elementary-school diagrams with manual region and relation annotations, all in English, with a strong bias toward natural-history and basic-physics content. ScienceQA reuses a similar source pool but pivots to QA pairs. None of these sources are stratified along finer content-area lines than the high-level subject, and none publish a frozen split. EduT2I is smaller and more curated: 1,685 figures, each anchored to the chapter it was extracted from and to a content-area tag.

Diagram code generation. A separate strand of work targets diagram *code* generation rather than pixel generation. AutomaTikZ (Belouadi et al., 2024) and DiagrammerGPT (Zala et al., 2023) treat the diagram-synthesis problem as a structured-output task and sidestep the rendering-quality questions that pixel-space generators face. We position EduT2I as complementary: the captions can be fed to either family of methods, and the rendered figures provide the ground-truth pixels needed to evaluate either.

Content-stratified splits and bootstrap uncertainty. The closest methodological reference for our split is the per-class stratification used in WebDataset benchmarks. The novelty here is the two-pass tagging step that pairs deterministic keyword rules with a local-LLM fallback. FID is biased upward at small sample sizes (Bińkowski et al., 2018; Chong and Forsyth, 2020); most text-to-image papers report point estimates at $n \geq 2{,}048$, but evaluation benchmarks at this scale are rare for specialised image domains. We report point FID and KID at $n{=}478$ alongside bootstrap 95% CIs on the per-sample metrics where bootstrap is tractable.

3 The EduT2I dataset

3.1 Source material

NCERT publishes the official secondary-school textbook series for the Indian curriculum. The full set covers 17 (class, subject) pairs: Science (single textbook each for classes 6–10), and Mathematics, Physics (Parts 1+2), Chemistry (Parts 1+2), and Biology for classes 11 and 12. We download the per-chapter PDFs directly from the NCERT distribution URLs under `ncert.nic.in/textbook/pdf/`. The class-6 Science textbook was re-released in 2024 under a new identifier (*Curiosity*). Chapters whose distribution URL returns an HTTP 404 response are excluded from the download loop, and the remaining 158 chapters are collected and indexed. We release only the build pipeline and the derived figure files. The source PDFs themselves are not redistributed.

3.2 Caption-anchored extraction

NCERT PDFs use a typesetting layer in which the `page.get_images()` call in PyMuPDF returns mostly page-mask and QR-code rasters rather than the illustrations a human reader sees — a behaviour also documented for other dense-typeset PDF corpora that motivated the custom extraction pipelines in DocLayNet (Pfitzmann et al., 2022) and PubLayNet (Zhong et al., 2019). Our extractor instead anchors on the caption text (Fig-

ure 1).

We scan each page for substrings matching `Fig\d` or `Figure\d`. For each match (Figure 1), the extractor takes the caption text block as the bottom anchor of the figure, walks upward in the same column until it hits the next text block, and saves everything between as the figure crop. Multi-column pages therefore yield one crop per column instead of one page-wide rectangle. This gives 1,685 figures across 158 chapters; 1,392 carry a caption, while 293 are saved without one because no caption could be matched nearby (per-split coverage rates are in Table 1).

A second pass cleans up the raw OCR captions with a local language model (gemma3:4b run via Ollama). The cleaned captions are written back into the dataset using `(lesson_globalID, fig_key)` as the lookup key, where `lesson_globalID` names the chapter (for example `L_NCERT_class10_science_05`) and `fig_key` names the figure within it. Because this key does not depend on the train/val/test assignment, the LLM cleanup step can be rerun at any time without redoing the figure extraction.

3.3 Content-area taxonomy

Every NCERT chapter covers one self-contained content area within its discipline: an optics chapter inside physics, a cell-biology chapter inside biology, a trigonometry chapter inside mathematics. We capture this structure with a flat 35-tag taxonomy that becomes the stratification key used in Section 3.5. Each chapter is assigned exactly one tag, and the label space is closed and committed once so that both the tagging script and the split assignment can rely on it.

The 35 tags partition into five disciplines. Representative tags include `optics` and `electromagnetism` (physics), `atomic_structure` and `organic` (chemistry), `cell_biology` and `genetics_heredity` (biology), and `calculus` and `coordinate_geometry` (mathematics). The full 35-tag list with discipline grouping is in Appendix A. The discipline grouping is for human readability only; the split algorithm treats the 35 tags as a flat label space.

3.4 Two-pass tagging

Assigning a taxonomy tag to a chapter is straightforward when the chapter's text mentions the tag's keywords directly (a chapter titled *Reflection of Light* is unambiguously `optics`) and harder when the chapter mixes content areas (a class-11 chemistry chapter on the periodic table can plausibly map to `atomic_structure`, `chemical_bonding`, or `inorganic_groups`). We use a two-pass system: a deterministic keyword pass handles the easy cases and a local-LLM fallback resolves the ambiguous ones. Both passes operate on a per-chapter text blob.

The blob is built from the lesson name, the headings of every *topic* subsection within the chapter (the textbook division that contains the figures we extracted), and the first ∼200 characters of each topic paragraph, truncated to 3,000 characters. We then run the blob through about 80 keyword rules. Each rule is a short list of words that, when any of them appears in the blob, votes for one taxonomy tag. Three abbreviated examples (each word list is anchored on word boundaries):

```
(reflection|lens|optic)   → optics
(electric|magnet|current) → electromagnetism
(orbital|aufbau)          → atomic_structure
```

The full rule list is in Appendix B.

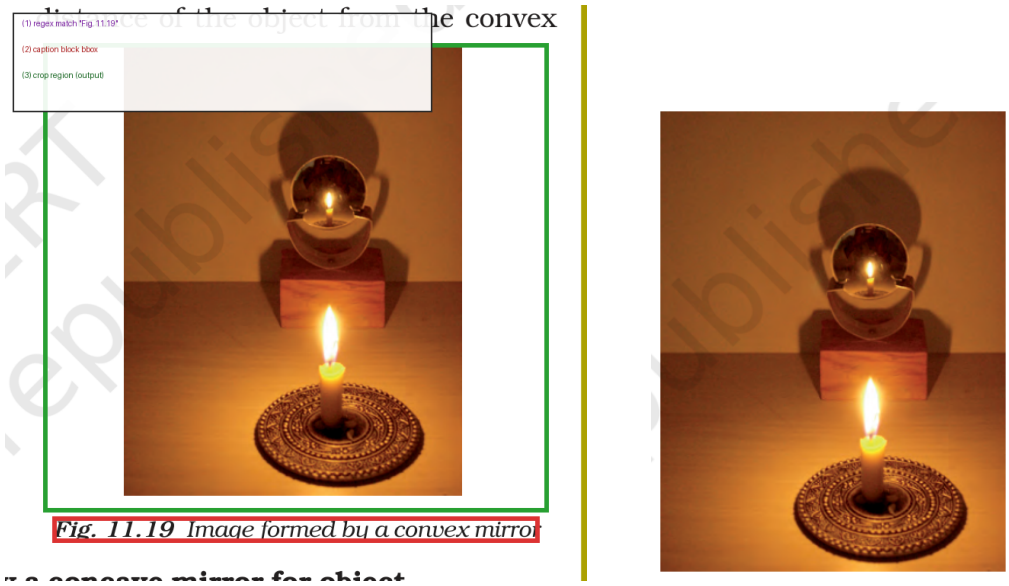
The tag whose rules accumulate the most hits wins the chapter. If no rule fires or the top two rules tie, the chapter is escalated to the LLM with the prompt "pick exactly one tag from $\{\dots\}$ that best describes this chapter. Reply with the tag only.". The LLM output is validated against the taxonomy and one retry is allowed on out-of-vocabulary returns. On the 158-chapter corpus, 144 chapters (91%) are tagged by the keyword pass and 14 (9%) by the LLM fallback; zero chapters fall through to a (class, subject) default. Two real LLM escalations are documented in Appendix B.

Of the 35 tags in the taxonomy, 32 are assigned to at least one chapter on the released corpus. The remaining three are present in the taxonomy for future extensibility. The five most-populated tags are `mechanics` (18), `electromagnetism` (13), `human_anatomy_physiology` (11), `states_of_matter` (10), and `evolution_diversity` (8).

3.5 Tag-stratified split

The split assignment is a deterministic seed-then-greedy algorithm parameterised on the global target ratios $(\text{train}, \text{val}, \text{test}) = (0.80, 0.10, 0.10)$. It processes one tag stratum at a time and runs in two phases.

Seed phase. Within each tag stratum we sort the



(a) source-page region with overlays **(b)** released figure

Figure 1: Worked example of the caption-anchored extractor on `gesc111.pdf`, page 9 (class 7 science, chapter 11, Light). **(a)** The regex pass matches the substring `Fig. 11.19` on the page; the text-block bounding box of the standalone caption "*Fig. 11.19 Image formed by a convex mirror*" (red) fixes the lower edge of the crop, while the preceding text block sets the upper edge and the column boundaries clamp it horizontally; the resulting crop region (green) is what the pipeline writes to disk. **(b)** The released figure, with its LLM-cleaned caption attached.

chapters by $(-n_{\text{figures}}, \text{globalID})$ — the heaviest chapter first, with the alphanumeric identifier as a tie-breaker so the ordering is fully determined by the input. The first three chapters in that order are seeded into train, val, and test respectively. Strata with fewer than three chapters seed only the splits they can fill (train only, or train + val).

Greedy phase. The remaining chapters in the stratum are assigned one at a time. Each chapter goes to the split whose current figure count is furthest below its tag-local target. If two splits are tied on the local deficit, the global deficit (the target ratios applied to the corpus-wide figure total) is the secondary tie-break.

Because every step is decided by deterministic ordering and counts — no RNG, no shuffle — the algorithm is stable: the same input `lesson_tags.json` always produces the same `splits.json`. The resulting split places 107 chapters into train, 29 into val, and 22 into test, with figure counts of 1,124 / 351 / 210 (Table 1).

Tag stratification carries one cost: 6 of the 32 active tags have fewer than three chapters, so the seed phase cannot place them in all three splits. These rare-tag chapters land in train only or train + val. We accept this because the two alternatives are both worse: over-sampling rare topics distorts the train distribution, and merging rare tags into bigger buckets defeats the content-stratification goal in the first place.

	Train	Val	Test	Total
Chapters	107	29	22	158
Figures	1,124	351	210	1,685
Figures w/ caption	918	272	202	1,392
Caption rate	81.7%	77.5%	96.2%	82.6%
Distinct tags	32	29	22	32

Table 1: Top-level dataset statistics. Tag stratification places every tag with ≥ 3 chapters into all three splits. Six rare tags (< 3 chapters) appear in train only or train + val only.

3.6 Schema and license

Each split ships as a JSON file in AI2D/TQA shape augmented with NCERT-specific provenance: per-chapter `globalID`, `lessonName`, `class/subject/chapter`, the assigned `tag`, and per-topic figure records with `caption` and `imagePath`. A worked example is in Appendix F. The image basename `<class>_<subject>_<chapter>_p<page>_<idx>.png` encodes the lesson identity so files can move between split folders without renaming when the split is regenerated.

NCERT publishes its textbooks for free educa-

tional reuse under the terms stated on the NCERT website (no redistribution fee, derivative works permitted for educational purposes). We re-release the extracted figures and the caption JSON under the same terms and do not redistribute the source PDFs. The corpus contains textbook illustrations rather than photographs of identifiable individuals, so the standard person-image-privacy concerns of natural-image datasets do not apply; users should nonetheless consult the NCERT terms before any deployment beyond research use.

4 Benchmark protocol

We evaluate five generators on EduT2I: three publicly available text-to-image diffusion models in a zero-shot setting (Stable Diffusion 1.5, SDXL-Turbo, and SDXL-base at 512), and two LoRA-fine-tuned variants of Stable Diffusion 1.5 trained on AI2D and on AI2D+ScienceQA respectively. All five share the same evaluation set, metric suite, and bootstrap-CI procedure.

Evaluation set. We evaluate on a content-stratified subset of 478 figures drawn from the EduT2I corpus. The sample is constructed by drawing 100 figures each from physics, chemistry, biology, and science, plus all 78 mathematics figures (the math pool is the smallest, so 78 is the ceiling, not a chosen cap). The sample is drawn from the union of train, val, and test, because none of the models under evaluation are fine-tuned on EduT2I — the two LoRA variants are trained on AI2D / AI2D+ScienceQA and the three zero-shot baselines never see any NCERT figure during training. For these models every NCERT figure is held out at train time, so the strict test-set-only argument does not apply. The sampler is seeded; same seed, same sample.

Metrics. For evaluating the five models we report six metrics on every run: FID (Heusel et al., 2017), KID (Bińkowski et al., 2018) with `subset_size=50`, LPIPS (Zhang et al., 2018) with the AlexNet backbone, SSIM, PSNR (both pure-torch implementations), and CLIPScore (Hessel et al., 2021) using a fixed `openai/clip-vit-base-patch32` backbone so cross-model comparisons share the same metric model. All generated images are resized to 512×512 before metric computation, regardless of the model's native generation resolution; ground-truth figures are resized on the same path.

Bootstrap confidence intervals. For LPIPS, SSIM, PSNR, and CLIPScore we save the per-sample value at evaluation time and resample 1,000 times with replacement from the n=478 vector. The reported 95% confidence interval (CI) is the $[2.5\%, 97.5\%]$ percentile of the bootstrap mean. FID and KID are not per-sample reducible without re-running Inception-v3 on every bootstrap sample, which is intractable at our compute budget; we report them as point estimates and acknowledge the small-n variance in Section 9.

5 Methods

5.1 Zero-shot baselines

Each baseline uses a fixed style prefix attached to the figure caption: "*educational diagram, clear scientific illustration, labeled, white background,*" followed by the caption. Models that accept negative prompts get a fixed negative prompt aimed at suppressing photographic and 3D-render output. All models run at fp16 on a single Apple Silicon MPS GPU.

SD 1.5 (runwayml/stable-diffusion-v1-5): 50 inference steps, guidance scale 7.5, generation at native 512. The same checkpoint underlies both LoRA variants.

SDXL-Turbo (stabilityai/sdxl-turbo): 4 inference steps, guidance scale 0 (the distilled model is trained to ignore guidance), generation at 512.

SDXL-base 1.0 (stabilityai/stable-diffusion-xl-base-1.0): 30 inference steps, guidance scale 7.5. We generate at 512×512 rather than native 1024×1024 to keep wallclock tractable on a 24 GB single-GPU budget; the gap that this introduces is noted in Section 9.

5.2 LoRA-fine-tuned pipeline

The two LoRA variants share an identical architecture and training recipe. Only the training corpus changes. The pipeline wraps Stable Diffusion 1.5 with three additions, shown in Figure 2.

Prompt path. The caption is fed through a BART summariser (`facebook/bart-large-cnn`) with maximum input 1,024 tokens and target summary 20–60 tokens. The summary becomes the prompt, so the 77-token CLIP cap is never hit on long captions. The same style prefix as the zero-shot evals is prepended.

LoRA adapter. We choose rank 8 as a balance between adapter expressiveness and param-

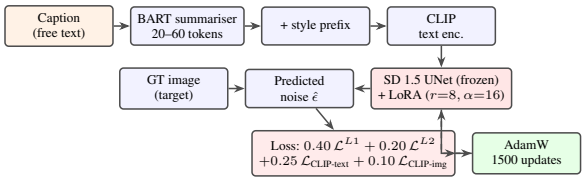


Figure 2: LoRA-fine-tuned pipeline (v3.5 recipe). The BART summariser collapses long captions to a 20–60-token prompt. Only the LoRA adapter on the UNet is trained. The four-term loss is fixed across both training corpora.

eter count on a 24 GB single-GPU budget; higher ranks (16, 32) trained more slowly without an obvious quality return in our small-scale ablations, and lower ranks (2, 4) under-fit the AI2D illustration style. With α=16 the effective LoRA scaling is $\alpha/r = 2$. Dropout is 0.05 and the adapter is applied to the UNet only; the text encoder and the VAE stay frozen.

Loss and schedule. A four-term loss with hand-tuned weights: $\mathcal{L} = 0.40\,\mathcal{L}^{L1}_{\text{recon}} + 0.20\,\mathcal{L}^{L2}_{\text{recon}} + 0.25\,\mathcal{L}_{\text{CLIP-text}} + 0.10\,\mathcal{L}_{\text{CLIP-img}}$. The reconstruction terms operate on the predicted noise. The CLIP-text term pulls the CLIP-text embedding of the generated image toward the CLIP-text embedding of the prompt; the CLIP-img term pulls the CLIP-image embedding of the generated image toward that of the ground truth. We use a cosine noise schedule with 1,000 training timesteps, 30 inference steps, and classifier-free guidance scale 8.5 at evaluation.

Optimiser. AdamW, learning rate 1×10^{-4}, 100 warm-up steps, batch size 1, gradient accumulation 4 (effective batch 4). Training runs for 6,000 forward steps (1,500 optimiser updates), seed 42. The combined-corpus LoRA sees a 50/50 mix of AI2D and ScienceQA; the AI2D-only LoRA sees the same number of updates on AI2D alone. Neither variant sees any NCERT figure during training.

Training takes $\sim$6 hours per LoRA on a single Apple M-series MPS GPU; zero-shot evaluation at n=478 takes $\sim$6 hours for the SD 1.5-class models, $\sim$1.5 hours for SDXL-Turbo (4 steps), and $\sim$5 hours for SDXL at 512. The full benchmark queue took $\sim$34 hours including model downloads.

6 Results

6.1 Aggregate metrics

Table 2 reports the headline numbers. SDXL-Turbo is the strongest performer on every metric except SSIM, where the LoRA variants edge ahead. The two fine-tuned models are the weakest on FID, KID, and CLIPScore; the bootstrap 95% CIs on LPIPS and CLIP do not overlap between the LoRA variants and the zero-shot baselines.

Three observations follow from Table 2.

Distillation does not hurt on this domain. SDXL-Turbo, the 4-step distilled variant, beats both its full-size parent SDXL-base (at 512) and the older SD 1.5. The LPIPS gap (0.748 vs. 0.830 vs. 0.778) and the PSNR gap (10.44 vs. 6.94 vs. 7.60) are wide with non-overlapping CIs. At zero-shot inference on out-of-distribution diagrams, the Turbo distillation step actually helps.

Fine-tuning regresses on FID. The combined-corpus LoRA is the worst FID at 201.5; the AI2D-only LoRA is second-worst at 184.4. Both are trained from the same SD 1.5 base whose zero-shot FID is 137.4, so the regressions are $+47$ and $+64$ FID respectively. The KID and CLIP regressions are smaller in magnitude but consistent in direction.

Mixing in ScienceQA makes things worse, not better. The combined LoRA loses to the AI2D-only variant on every metric: $+17$ FID, $+0.011$ KID, $+0.015$ LPIPS, -0.005 SSIM, -0.001 CLIP. The mechanism behind this is discussed in Section 7.1.

6.2 Per-subject breakdown

Table 3 reports the per-subject FID and a Δ FID column quantifying how much the AI2D-only variant beats the combined one per subject. The AI2D-only LoRA wins on every subject; the gap is largest on math (-36 FID) and smallest on science (-1). The physics and biology subjects sit closest to AI2D's natural-history visual idiom (FID 251 and 258 respectively for the AI2D-only LoRA) and benefit the most from the AI2D-only training mix; chemistry and math sit furthest (FID 281 each).

6.3 Qualitative analysis

We show one representative panel per subject in Figure 3 (additional samples in Appendix G). Panels were chosen by ranking samples on their combined per-sample LPIPS so that the comparison reflects each model at or near its best output on the caption, not a worst-case crash. In every panel the ground-truth figure is on the left and the generated image is on the right.

Model (n=478)	**FID** ↓	**KID** ↓	**LPIPS** 95% CI ↓	**SSIM** 95% CI ↑	**PSNR** 95% CI ↑	**CLIP** 95% CI ↑
SDXL-Turbo	**133.73**	**0.044**	**0.748 [0.744, 0.753]**	0.051 [0.046, 0.057]	**10.44 [10.26, 10.62]**	0.276 [0.274, 0.279]
SD 1.5 (zero-shot)	137.37	0.054	0.830 [0.825, 0.835]	0.030 [0.024, 0.036]	6.94 [6.81, 7.07]	0.275 [0.272, 0.278]
SDXL @ 512	149.26	0.079	0.778 [0.774, 0.783]	0.025 [0.020, 0.030]	7.60 [7.44, 7.77]	0.276 [0.273, 0.279]
LoRA, AI2D only	184.38	0.083	0.822 [0.816, 0.828]	0.021 [0.018, 0.025]	5.56 [5.46, 5.67]	0.251 [0.250, 0.253]
LoRA, AI2D + ScienceQA	201.48	0.094	0.837 [0.831, 0.844]	0.016 [0.013, 0.020]	5.56 [5.46, 5.67]	0.250 [0.249, 0.252]

Table 2: Main benchmark results at n=478 stratified across five subjects. Bootstrap 95% CI on LPIPS, SSIM, PSNR, and CLIPScore at 1,000 resamples. FID and KID are point estimates.

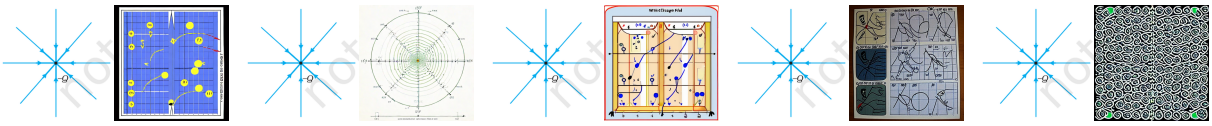


Figure 3: Physics — electric field around a point charge (sample 41, caption: "*Electric field (a) due to a charge Q, (b) due to a charge −Q.*"). In every panel the ground-truth (GT) figure is on the left and the generated image is on the right. SDXL-Turbo (best LPIPS at 0.582) is closest to the GT's flat-field illustration. The LoRA variants flatten the colour palette but lose the radial-field structure.

Subject (n)	**Combined**	**AI2D-only**	**Δ FID**
biology (100)	272.76	257.68	−15.1
chemistry (100)	293.44	281.38	−12.1
math (78)	316.50	280.54	−36.0
physics (100)	279.57	251.32	−28.3
science (100)	267.44	266.05	− 1.4
aggregate (478)	201.48	184.38	−17.1

Table 3: Per-subject FID for the two LoRA variants. The Δ FID column is (AI2D-only − Combined). Aggregate FID is markedly lower than every single-subject FID because the aggregate set is more diverse (see Section 7.1). Per-subject CLIPScores are in Appendix C.

Three qualitative patterns dominate, all visible in Figure 3 and the appendix samples. *First*, every model defaults to realistic-looking content even when the GT is a flat schematic; SD 1.5 produces photographically textured output, SDXL-Turbo flattens texture but adds background detail, and the LoRA variants pull toward AI2D-style flat illustration but lose entity fidelity. *Second*, none of the models render text labels on the diagrams reliably — the CLIP prior cannot render arrows-and-labels, so every model substitutes textured fill for what should be annotated structure. This is the single largest gap between the generated and ground-truth distributions and drives most of the LPIPS and FID separation. *Third*, the LoRA variants exhibit *mode collapse toward training-distribution style*: similar prompts produce visually similar outputs because the AI2D-induced prior overwhelms the caption signal — the most likely mechanism behind the FID regression.

7 Discussion

7.1 Why LoRA fine-tuning hurts

We see three plausible mechanisms behind the LoRA regression and the data is most consistent with the third.

(i) Insufficient training. 1,500 optimiser updates is short. But the AI2D-only LoRA is uniformly better than the combined one even though both saw the same number of updates, and the combined regression is *larger* than the AI2D-only one — the opposite of what longer-training would predict if more data helped.

(ii) Wrong loss weights. The hand-tuned four-term loss is a contributing factor but not dominant: both LoRAs share the same weights and the combined LoRA loses to the AI2D-only one anyway.

(iii) Distribution shift between training corpus and target. AI2D diagrams are small, mid-school, mostly biology and physics, in a specific cartoony style. ScienceQA contributes more diverse but photographically-grounded content. NCERT has its own visual idiom: denser labels, skeletal-chemistry notation, typeset-math overlays, textbook-grade rendering. Fine-tuning on AI2D pulls the SD 1.5 prior toward AI2D-style flat illustration; adding ScienceQA pulls it in a third direction. Neither pull points at NCERT, and the combination of two distinct pulls is the worst of both worlds. The per-subject breakdown supports this: the regression is largest on math (no AI2D

analogue) and smallest on physics and biology (where AI2D has the most coverage).

We further confirm in Appendix D that the same combined LoRA produces clean, diagrammatic output when evaluated *in-domain* on AI2D (FID 240.97) and ScienceQA (FID 248.17), materially lower than any of its per-subject NCERT FIDs. The training procedure is not broken — the bottleneck is cross-domain transfer to NCERT.

The practical implication is that "train a LoRA on the nearest diagram corpus" is not a reliable transfer recipe. EduT2I makes this kind of negative finding quantitative and reproducible.

7.2 Why SDXL-Turbo wins

The Turbo distillation procedure (Sauer et al., 2023) optimises for high-quality output at low step count, trading FID for perceptual quality on natural images. On educational diagrams the trade is favourable: the simpler, lower-frequency outputs Turbo produces match the simpler, lower-frequency targets in NCERT better than the high-frequency texture that full SDXL adds. SDXL @ 1024 would presumably do better on FID and worse on wallclock; whether that beats Turbo is the most natural follow-up the benchmark invites.

8 Conclusion

EduT2I is a 1,685-figure, content-stratified, caption-aligned benchmark for educational diagram synthesis. Five publicly available text-to-image generators — including two we LoRA-fine-tune ourselves on the closest neighbour corpora — all fail to reach a usable quality bar on Indian secondary-school content. The fine-tuning step *regresses* every metric versus the zero-shot model it was built on, and the regression is larger when the fine-tuning corpus mixes AI2D with ScienceQA than when it uses AI2D alone. We read this as evidence that distribution shift between the LoRA training corpus and the target benchmark matters more than the size of the training corpus, at least at the rank-8, 1,500-update scale we tested. The dataset, the splits, the evaluation harness, and 2,390 generated panels are released so that future work can target the specific failure modes we document.

9 Limitations

FID variance at n=478. FID is biased and high-variance below $n \approx 2{,}048$. Our sample is below that threshold. We mitigate by reporting KID, which is less biased at small n, and by reporting bootstrap CIs on the per-sample metrics. The FID and KID rankings agree, which gives some confidence that the model ordering is robust, but absolute FID numbers should be read as in the range of ∼140–200 rather than as precise point estimates. The bootstrap CIs we do report cover only the per-sample metrics (LPIPS, SSIM, PSNR, CLIPScore) because per-bootstrap FID re-computation requires re-running Inception-v3, which is intractable at our compute budget.

Single-architecture coverage on the LoRA side. Both fine-tuned variants are SD 1.5 + LoRA. We did not run LoRA on top of SDXL, SDXL-Turbo, or any of the more recent MMDiT-based models. The negative-result framing applies to the SD 1.5 family; whether the same pattern holds for SDXL-class adapters is an open question that the benchmark and the released training pipeline support.

SDXL evaluated at 512, not 1024. SDXL is trained for 1024×1024 generation. We evaluate at 512 to keep wallclock tractable on the 24 GB single-GPU budget. The reported metrics are therefore an under-estimate of what SDXL is capable of on this corpus.

No human evaluation. We do not include human ratings of "usable as an educational diagram", which would be the most reliable measure of the third property the benchmark targets (structured, labelled output). We omitted it because of budget, not because it is uninformative; the 2,390 generated panels are released so that a follow-up study can run such ratings cleanly.

Single curriculum. The dataset is drawn from one national textbook series. The framing as "out-of-distribution for AI2D" is specific to that series; non-Indian non-Western curricula may behave differently. Extending the pipeline to additional national curricula is straightforward given the build scripts and the frozen taxonomy.

Compute constraints. All experiments ran on a single 24 GB Apple Silicon MPS GPU with fp16. Several baselines we wanted to evaluate (FLUX.1-dev at 12B parameters; gated repositories requiring an authenticated HuggingFace token in our deployment environment) could not be run on this

hardware and are absent from Table 2. The protocol supports them; the evaluation is not yet exhaustive.

10 Ethical considerations

Source licensing. The corpus is derived from NCERT textbooks, which are publicly distributed by the Indian government for educational reuse. We re-release the extracted figures and captions under the same terms; we do not redistribute the source PDFs. Users targeting non-research applications should consult the NCERT licence directly.

Personally identifying information. The corpus contains textbook illustrations (diagrams, schematics, line drawings) rather than photographs of identifiable individuals. The standard person-image-privacy concerns of natural-image datasets do not apply to this corpus.

Cultural representation. The dataset is drawn from one national curriculum (Indian secondary-school NCERT) and so does not represent the diagram conventions or scientific notation used in other curricula. Training generators exclusively on this corpus would embed Indian-textbook conventions; users mixing it with other corpora should be explicit about which conventions they expect their generator to produce.

Negative-result framing. The headline finding is that current text-to-image fine-tuning recipes do not produce usable educational diagrams in this domain. We release the benchmark expecting it to be used to measure progress, not to claim that the tested methods are themselves ready for classroom deployment.

Use of generative AI. Generative AI was used in preparing this submission for two purposes: (i) LaTeX writing assistance, including drafting, formatting, and prose polishing of the manuscript, and (ii) assistance organising the evaluation harness and figure preparation scripts. All experimental results, dataset construction decisions, and analyses are the authors' own; the AI assistance was used in the language and code-presentation roles described in the ACL Policy on AI Writing Assistance, categories (a) and (d). All text and code were reviewed and verified by the authors.

References

Jonas Belouadi, Anne Lauscher, and Steffen Eger. 2024. AutomaTikZ: Text-guided synthesis of scientific vector graphics with TikZ. ArXiv preprint, ICLR 2024.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. 2018. Demystifying mmd gans. In *International Conference on Learning Representations (ICLR)*.

Min Jin Chong and David Forsyth. 2020. Effectively unbiased fid and inception score and where to find them. In *CVPR*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*.

Tuomo Hiippala, Malihe Alikhani, Jonas Haverinen, Timo Kalliokoski, Evanfiya Logacheva, Serafina Orekhova, Aino Tuomainen, Matthew Stone, and John A. Bateman. 2021. Ai2d-rst: A multimodal corpus of 1000 primary school science diagrams. *Language Resources and Evaluation*.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *European Conference on Computer Vision (ECCV)*.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S. Nassar, and Peter Staar. 2022. DocLayNet: A large human-annotated dataset for document-layout analysis. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. 2023. Adversarial diffusion distillation. In *arXiv preprint arXiv:2311.17042*.

Abhay Zala, Jaemin Cho, and Mohit Bansal. 2023. DiagrammerGPT: Generating open-domain, open-platform diagrams via LLM planning. ArXiv preprint arXiv:2310.12128.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*.

Xu Zhong, Jianbin Tang, and Antonio Jimeno-Yepes. 2019. PubLayNet: Largest dataset ever for document layout analysis. In *International Conference on Document Analysis and Recognition (ICDAR)*.

A Full 35-tag taxonomy

The taxonomy is shipped as a frozen JSON. Tags are grouped under five disciplines for human readability; the split algorithm treats them as a flat label space.

- **physics** (7): mechanics, thermodynamics, optics, electromagnetism, waves_oscillations, modern_physics, fluids_matter
- **chemistry** (8): atomic_structure, chemical_bonding, thermochemistry, equilibrium_acids_bases, redox_electrochem, organic, inorganic_groups, states_of_matter
- **biology** (8): cell_biology, genetics_heredity, evolution_diversity, ecology_environment, human_anatomy_physiology, plant_biology, microbiology_disease, biotechnology_reproduction
- **math** (8): algebra, calculus, coordinate_geometry, trigonometry, probability_statistics, matrices_determinants, sets_relations_functions, vectors_3d
- **general** (4): physics_intro, chemistry_intro, biology_intro, earth_environmental

The top-10 most populated tags on the released corpus are `mechanics` (18 chapters), `electromagnetism` (13), `human_anatomy_physiology` (11), `states_of_matter` (10), `evolution_diversity` (8), `cell_biology` (8), `organic` (7), `biotechnology_reproduction` (7), `sets_relations_functions` (6), and `genetics_heredity` (5).

B Two-pass tagging: control flow and worked LLM escalations

LLM escalation 1. Class-11 chemistry part-2 chapter 3 triggers seven candidate tags with only 1–2 keyword hits each (fluids_matter 1, atomic_structure 2, chemical_bonding 1, redox_electrochem 2, organic 2, algebra 1, sets_relations_functions 1). With no clear winner the chapter is routed to the LLM, which resolves it to organic from the topic headings (alkanes, alkenes).

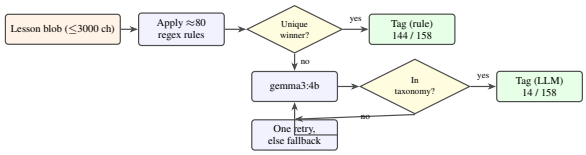


Figure 4: Two-pass tagging control flow. The keyword pass resolves 91% of chapters unambiguously; the remaining 9% escalate to a local LLM (gemma3:4b via Ollama) whose output is validated against the frozen taxonomy. Zero chapters fell through to a (class, subject) fallback on the released corpus.

LLM escalation 2. Class-11 mathematics chapter 1 accumulates only two stray hits from physics rules (states_of_matter 1, earth_environmental 1) because the chapter's vocabulary does not overlap any math keyword rule; the LLM picks sets_relations_functions from the title.

C Per-subject metrics for both LoRA variants

Subj. (n)	**Combined**			**AI2D-only**		
	FID	LPIPS	CLIP	FID	LPIPS	CLIP
biology (100)	272.76	0.838	0.240	257.68	0.810	0.241
chemistry (100)	293.44	0.842	0.254	281.38	0.830	0.257
math (78)	316.50	0.828	0.255	280.54	0.814	0.259
physics (100)	279.57	0.857	0.255	251.32	0.840	0.255
science (100)	267.44	0.821	0.248	266.05	0.814	0.247
aggregate (478)	201.48	0.838	0.250	184.38	0.822	0.251

Table 4: Per-subject FID, LPIPS, and CLIP for both LoRA variants.

D In-domain capability check

To rule out the explanation that the LoRA recipe is broken, we evaluate the combined-corpus LoRA on the corpora it was trained on (AI2D and ScienceQA) and on NCERT for reference (Table 5). In-domain FID is 240.97 on AI2D and 248.17 on ScienceQA, both materially lower than any per-subject NCERT FID of the same model (251–316). The same v2→v3.5 recipe transition (enabling CLIP loss, switching to a cosine schedule, dropping the L4 perceptual term) drops in-domain AI2D FID from 317.74 to 240.97 — a 24% reduction that demonstrates the training procedure

works as intended. The bottleneck is cross-domain transfer to NCERT, not the recipe itself.

Group	Setting	n	FID	CLIP
v3.5 LoRA in-domain	AI2D	128	240.97	0.223
v3.5 LoRA in-domain	ScienceQA	128	248.17	0.224
v2 LoRA in-domain	AI2D	64	317.74	0.227
v2 LoRA in-domain	ScienceQA	64	256.30	0.225
v3.5, NCERT (worst)	math	78	316.50	0.255
v3.5, NCERT (best)	science	100	267.44	0.248
v3.5, NCERT	aggregate	478	201.48	0.250

Table 5: In-domain vs out-of-domain performance for the combined LoRA. Aggregate NCERT FID is lower than any single-subject NCERT FID and lower than the in-domain numbers because Inception's feature-distribution distance benefits from sample diversity.

E Hyperparameters

	SD 1.5	Turbo	SDXL	LoRA-c	LoRA-a
inf. steps	50	4	30	30	30
cfg	7.5	0.0	7.5	8.5	8.5
schedule	lin	def	def	cos	cos
gen size	512	512	512	512	512
neg prompt	y	n	y	y	y
LoRA rank	—	—	—	8	8
LoRA α	—	—	—	16	16
dropout	—	—	—	0.05	0.05
opt updates	—	—	—	1500	1500
batch / accum	—	—	—	1 / 4	1 / 4
lr	—	—	—	1e-4	1e-4
training corpus	—	—	—	AI2D+SciQA	AI2D

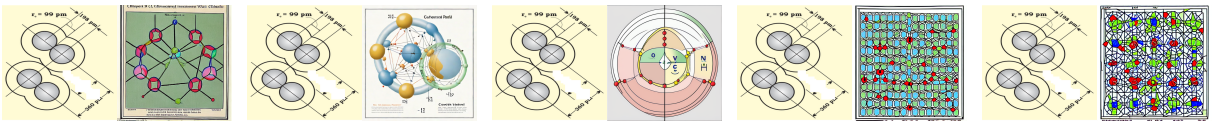
Table 6: Hyperparameters for the five evaluated models. "LoRA-c"= combined-corpus LoRA, "LoRA-a"= AI2D-only LoRA. "SDXL" is SDXL-base evaluated at 512×512.

F Schema sample


```
{
 "globalID": "L_NCERT_class10_science_05",
 "lessonName":
   "Human Eye and the Colourful World",
 "class": "10", "subject": "science",
 "chapter": 5, "tag": "optics",
 "topics": {"T_01": {
   "topicName":
     "Power of Accommodation",
   "figures": [{
     "globalID":
       "L_NCERT_class10_science_05_F_03",
     "caption":
       "Figure 11.3 The defect of myopia ...",
     "imagePath":
       "test/textbook_images/...p157_03.png"
   }]
 }}
}
```


G Additional sample panels

Each row is one NCERT sample across all five models. Indices were chosen by per-subject LPIPS rank with a caption-length filter (≥ 35 characters of semantic content). In every panel the ground-truth figure is on the left and the generated image is on the right; captions are abbreviated.

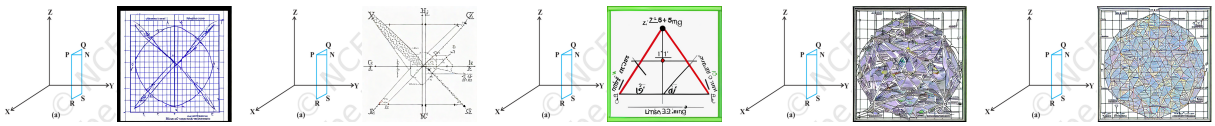


SD 1.5 Turbo SDXL AI2D LoRA Comb. LoRA

Figure 5: Chemistry — covalent and van der Waals radii in a chlorine molecule. Ground truth is on the left of each panel, generated image on the right.

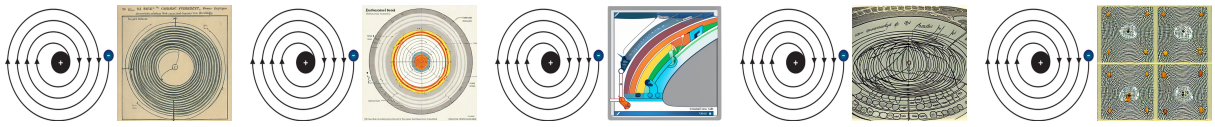
SD 1.5 Turbo SDXL AI2D LoRA Comb. LoRA

Figure 6: Biology — heterophylly in larkspur and buttercup. Ground truth is on the left of each panel, generated image on the right.



SD 1.5 Turbo SDXL AI2D LoRA Comb. LoRA

Figure 7: Mathematics — direction cosines of a line in 3D. Ground truth is on the left of each panel, generated image on the right.



SD 1.5 Turbo SDXL AI2D LoRA Comb. LoRA

Figure 8: Science — spiral path of a charged particle losing energy. Ground truth is on the left of each panel, generated image on the right.

Figure 9: Combined LoRA evaluated in-domain on AI2D (ground truth on the left of each panel, generated image on the right). Visibly more diagrammatic, more labelled, and more entity-faithful than the LoRA's NCERT outputs in Figure 3 — direct visual confirmation that the recipe trains a competent diagram generator within its training distribution.