FURLA: Furniture Layout Generation in Irregular Rooms with Human-Aligned Rewards

Anonymous ACL submission

Abstract

Real-world furniture layout is harder than typical benchmarks suggest: rooms have non-convex boundaries, doors and windows are fixed by building norms before placement, and no structured ergonomic reward guides learning. We present **FURLA**, a system tackling these issues jointly. Contributions include a dataset of 53,332 rooms from 22,000 real apartments with irregular polygons and pre-positioned openings, an eight-criterion expert-annotated ergonomic rubric, and a LoRA-tuned VLM critic that converts expert scores into a reinforcement-learning reward. The three-stage pipeline—SFT generator, LoRA-fine-tuned critic, and GRPO reward optimisation—is the first to close the loop between geometric validity and human-aligned ergonomic quality in a single trainable system. Using full polygon vertices instead of axis-aligned boxes enables exact wall-contact and door-clearance reasoning, matching constraints that designers follow in practice. Supervised fine-tuning cuts out-of-boundary and overlap rates by 74 % and 55 % versus zero-shot; the trained critic matches human inter-annotator agreement on ergonomic scoring. GRPO adds a further 1.4 % gain in mean ergonomic reward on held-out rooms.

1 Introduction

Placing furniture inside a room requires simultaneously respecting irregular wall boundaries, preserving door and window clearances, maintaining walkable circulation paths, creating coherent activity zones, and aligning seating with focal elements. Despite impressive visual and language capabilities, general-purpose vision-language models (VLMs) routinely fail at this task when room boundaries are non-convex—the common case in real apartments (Feng et al., 2023; Yang et al., 2024, 2025). This failure has two root causes: pre-trained models receive no geometry-specific supervision for furniture layout, and existing evaluation pipelines rely on simple collision metrics or pairwise preferences without a structured multi-criterion ergonomic rubric. In real architectural workflows, door and window positions are fixed by structural constraints, building codes, and fire-egress requirements before any furniture is placed, yet most prior work operates on abstract rectangular rooms with no openings (Paschalidou et al., 2021; Feng et al., 2023; Yang et al., 2024).

To address these challenges, we introduce **FURLA—Fur**niture **La**yout Generation in Irregular Rooms with Human-Aligned Rewards—a three-stage pipeline (Figure 2) that integrates explicit geometric reasoning with a learned ergonomic reward function.

Stage 1: Layout generator. We collect 22,000 real apartment floor plans (53,332 rooms, 13 types, 45 furniture categories) with non-convex boundaries and pre-positioned doors and windows. A VLM is fine-tuned via SFT to output full polygon vertex coordinates per furniture item, enabling exact wall-contact and door-clearance reasoning.

Stage 2: Ergonomic critic. We introduce an eight-criterion rubric (Table 1) covering functional zones, focal points, walkways, accessibility, furniture pairings, spatial balance, natural light, and entry visibility. Two experts annotate 1,121 floor plans; a LoRA-tuned VLM critic learns to predict per-criterion scores, reaching substantial expert agreement.

Stage 3: Reinforcement learning. The SFT generator and critic are connected in a GRPO loop: the policy samples four candidate layouts per prompt, the critic scores each on all eight criteria, and policy gradients are computed from within-group advantages, improving ergonomic reward over the SFT baseline.

Our contributions are: a 53,332-room polygon-vertex layout dataset—the first large-scale resource of this kind; an eight-criterion ergonomic rubric

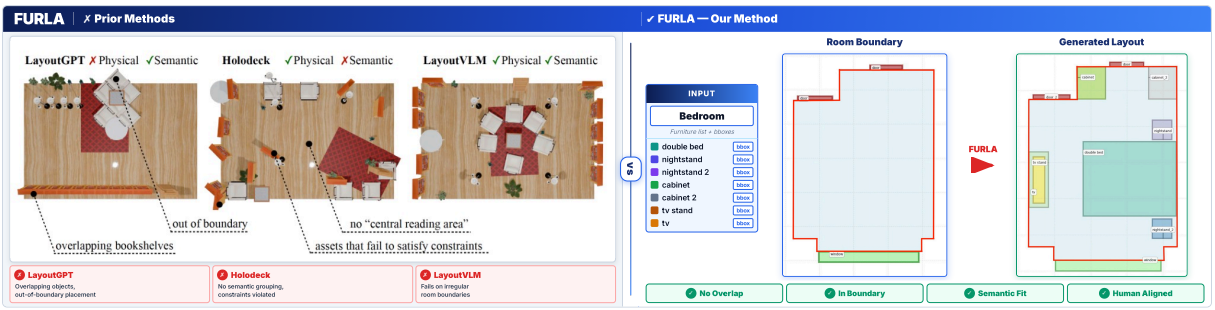


Figure 1: **FURLA compared to prior methods.** While LayoutGPT, Holodeck, and LayoutVLM produce physically invalid or semantically incoherent layouts in irregular rooms, FURLA generates arrangements that are overlap-free, within boundaries, semantically coherent, and aligned with ergonomic criteria.

with expert validation; a LoRA-tuned VLM critic evaluated across 36 configurations; and the FURLA pipeline demonstrating consistent ergonomic improvement over SFT alone. Sections 2–6 detail each component in turn.

2 Datasets and Problem Statement

FURLA uses three distinct datasets, one per stage. Dataset statistics and annotation quality are in Appendix A (Figures 6 and 8). Examples of floor plans are shown in Figure 3 below.

2.1 Stage 1 Dataset: SFT Layout Generation

We collected publicly available apartment floor plans from real-estate developer websites spanning studio to five-room layouts. Room boundaries were traced as polygon chains and every furniture piece labelled with its category and placed polygon vertices (examples in Figure 3). The vocabulary covers 45 furniture categories across 13 room types (full record structure in Appendix A.1). The defining characteristic is the irregular, non-convex polygonal boundary—substantially harder than axis-aligned rooms assumed by prior methods (Feng et al., 2023; Yang et al., 2024, 2025; Ran et al., 2025). The 53,332 rooms are split 70/15/15 (train 37,330 / val 8,001 / test 8,001).

2.2 Stage 2 Dataset: Critic Annotation

We designed an eight-criterion rubric grounded in ergonomic principles (Table 1; full details in Appendix B). Each criterion uses an ordinal scale (1–2, 1–3, or 1–4); higher is better. Priority tiers map to GRPO reward weights: CRITICAL (w=2) > HIGH (w=1.5) > IMPORTANT (w=1). C1, C2, C4–C6 are skipped for bathrooms/corridors/hallways/wardrobes; C7 requires a window; C8 requires a door.

Two expert annotators independently scored 1,121 floor-plan images (800×600 px, 2D top-down renders) across all applicable criteria, yielding 20,160 criterion–image pairs (80.5 % applicable). The dataset is split 70/15/15: per-criterion training yields 14,112/3,024/3,024 records; the aggregated strategy (all criteria in one call) yields 1,508/366/368 records (Appendix A.2).

Inter-annotator agreement. Cohen's weighted kappa (Cohen, 1968) gives κ=0.515 (moderate (Landis and Koch, 1977)), 66.2 % exact match, and 92.3 % adjacent-or-exact—the human quality ceiling for automated critique.

2.3 Stage 3 Dataset: GRPO Training Prompts

Each GRPO prompt encodes room type, polygon vertices, canonical furniture templates, door/window bounding boxes, and the SFT-initialised layout to be refined (prompt format in Appendix D.4). The policy samples K=4 candidate layouts per prompt, each scored by the critic on all eight criteria. The dataset contains 800 training and 200 validation prompts (3,200 and 800 candidates; Appendix A.2).

2.4 Problem Statement

Having described the three datasets, we now formalise the task each stage solves. Let a room $\mathcal{R}$ be defined by boundary polygon $\mathbf{P}^{\text{room}}=\{v_i\}_{i=1}^{N}$, door bounding boxes $\{D_j\}$, window bounding boxes $\{W_k\}$, type τ, and canonical furniture templates $\{F_l^0\}_{l=1}^{M}$. **Stage 1** finds generator G_θ:

$$G_\theta(\tau, \mathbf{P}^{\text{room}}, \{D_j\}, \{W_k\}, \{F_l^0\}) \longmapsto \hat{\mathcal{L}} = \{(\mathbf{P}_l^{\text{placed}}, o_l)\}_{l=1}^{M}, \quad (1)$$

s.t. placed polygons inside $\mathbf{P}^{\text{room}}$, non-exempt items non-overlapping, doors/windows unoccluded.

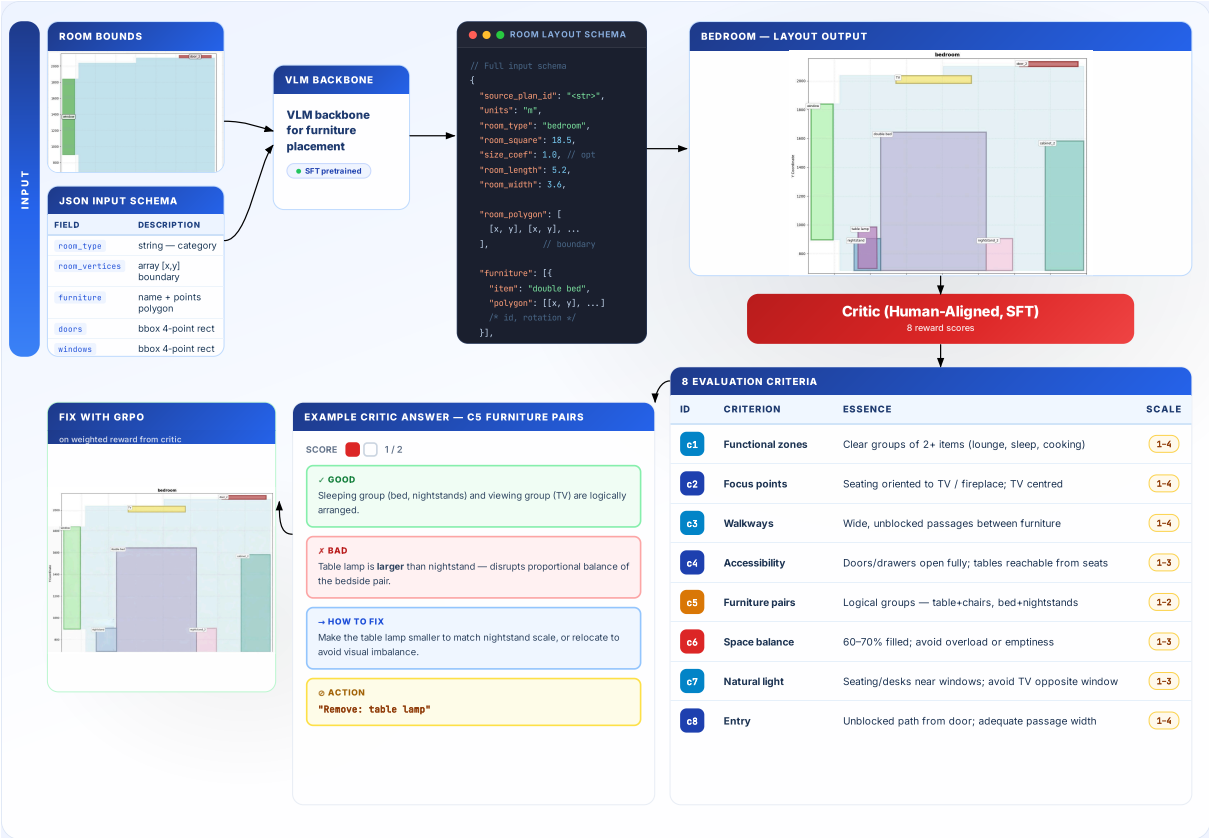


Figure 2: **FURLA pipeline.** A polygon-room prompt feeds an SFT-trained generator (Qwen3-VL-8B), whose candidate layouts are scored by a fine-tuned VLM critic (Qwen3-VL-32B) on eight ergonomic criteria. GRPO closes the loop, updating the generator from within-group critic rewards.

Stage 2 finds critic C_ϕ:

$$C_\phi(\mathcal{I}(\hat{\mathcal{L}}), \tau) \longmapsto \mathbf{s} = (s_1, \ldots, s_8), \quad s_c \in \{1, \ldots, K_c\}, \quad (2)$$

where $K_c \in \{2, 3, 4\}$. **Stage 3** maximises aggregate reward:

$$\max_\theta \; \mathbb{E}_{\hat{\mathcal{L}} \sim G_\theta}[r(\hat{\mathcal{L}}, \tau)] - \beta \cdot \mathrm{KL}[G_\theta \,\|\, G_{\theta_{\mathrm{SFT}}}], \quad (3)$$

where

$$r(\hat{\mathcal{L}}, \tau) = \frac{\sum_{c \in \mathcal{A}(\tau)} w_c (s_c - 1)}{\sum_{c \in \mathcal{A}(\tau)} w_c (K_c - 1)} \in [0, 1], \quad (4)$$

weights $w_c \in \{2, 1.5, 1\}$ follow priority tiers, and β anchors the KL divergence from the SFT policy.

3 Method

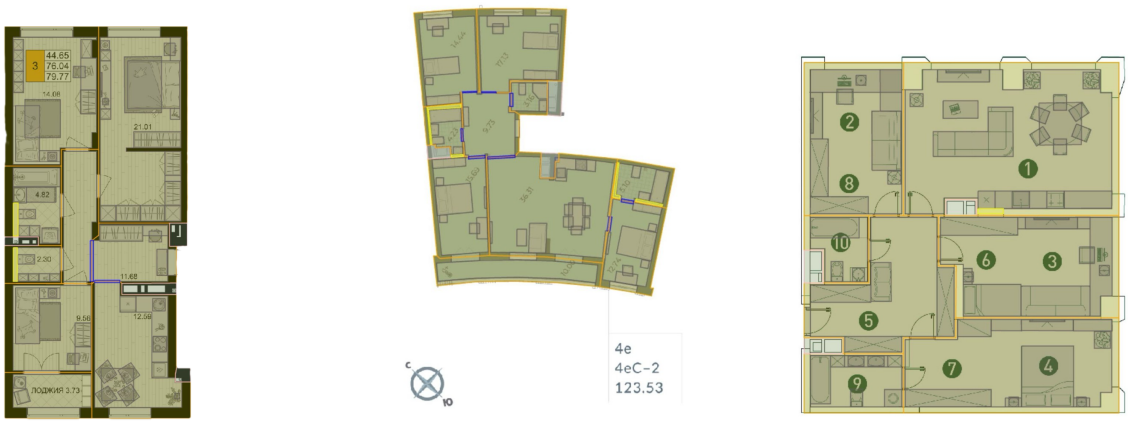
The FURLA pipeline (Figure 2) has three stages: SFT layout generation, VLM critic training, and GRPO optimisation. Each stage builds directly on the datasets described in Section 2; full hyperparameters and metric definitions are provided in Appendices C and B.

3.1 Stage 1: SFT Furniture Layout Generator

We initialize from `Qwen3-VL-8B-Instruct` (Bai et al., 2025a). Each training example provides a rasterized floor-plan image (walls, doors, windows) and a structured text payload containing the room polygon, canonical furniture polygons (longest side horizontal, bottom-left at $(0, 0)$), door/window bounding boxes, and room type. The model predicts rotation and translation for each item and outputs a JSON array of object–polygon pairs; doors and windows are hard-constraint inputs and excluded from output. LoRA adapters (Hu et al., 2022) (r=32, α=64, all Q/K/V/MLP modules) are trained with AdamW, learning rate 2×10^{-4}, cosine schedule, 10 % warmup, 15 epochs on 8×H100-80 GB with DeepSpeed ZeRO-3 (Rasley et al., 2020) and bfloat16. Full hyperparameters are in Appendix C.

3.2 Stage 2: VLM Critic Training

The critic backbone is `Qwen3-VL-32B` (Bai et al., 2025a) with LoRA (Hu et al., 2022). It is trained to score a rendered layout image against the eight



(a) Labeled example 1 (b) Labeled example 2 (c) Labeled example 3

Figure 3: **Dataset examples (Stage 1).** Human-annotated flat segmentation with 13 room types and 45 furniture class objects.

Table 1: **Eight ergonomic criteria (Stage 2).** Grouped by priority: CRITICAL (w=2) / HIGH (w=1.5) / IMPORTANT (w=1).

ID	Criterion	Scale	w	Description
C3	Walkways	1–4	2.0	Unobstructed circulation
C4	Accessibility	1–3	2.0	Cabinet doors open freely
C8	Entry Visibility	1–4	2.0	Clear passage from entrance
C1	Functional Zones	1–4	1.5	Distinct activity areas
C2	Focus Points	1–4	1.5	Seating faces focal elements
C5	Furniture Pairs	1–2	1.5	Related items grouped
C6	Spatial Balance	1–3	1.0	60–70 % fill, no dead zones
C7	Natural Light	1–3	1.0	Key zones lit, no glare

ergonomic criteria (Table 1); full prompt templates are in Appendix D.

We compare two distinct fine-tuning and inference strategies:

Per-criterion evaluates each of the eight criteria in a separate inference call—one prompt per criterion, one integer score as output. This yields ≈8 API calls per layout but isolates each criterion so the model can focus its attention and produce a more reliable score. The training set contains 14,112 annotated criterion–image pairs.

Aggregated evaluates all eight criteria simultaneously in a single call—one prompt outputs a JSON object with all eight scores and comments. This requires only 1× the inference cost but forces the model to share attention across all criteria at once, which trades accuracy for efficiency. The training set contains 1,508 annotated images.

We report MAE, Quadratic Weighted Kappa (QWK) (Cohen, 1968), Spearman ρ, METEOR (Banerjee and Lavie, 2005), and Cohen's d (Cohen, 1988) (definitions in Appendix B).

3.3 Stage 3: GRPO Joint Optimization

Stage 3 connects the SFT generator as policy with the critic as reward model in an online RL loop (Figure 4). Only LoRA parameters are updated via GRPO (Shao et al., 2024).

At each step: sample a prompt, generate K=4 JSON layouts, validate geometry, apply structural penalties (p_{ov}=0.05 per overlap, p_{bd}=0.08 per boundary violation), render valid layouts, score with the critic, compute within-group advantages $A^{(k)}=r^{(k)}-\bar{r}$, and update via clipped-ratio GRPO with KL penalty β=0.01 (Equation (3)). Four critic HTTP services (one per GPU, two criteria each) handle concurrent scoring with exponential-backoff retries (Appendix C.4). Full algorithm is in Appendix E.

4 Results

We evaluate each pipeline stage on its dedicated test set, then report the end-to-end ergonomic gain delivered by the GRPO loop. Stage 1 focuses on geometric quality (boundary/overlap violations and GT alignment); Stage 2 on critic agreement with expert annotations across two inference strategies; Stage 3 on per-criterion and aggregate reward gain after GRPO.

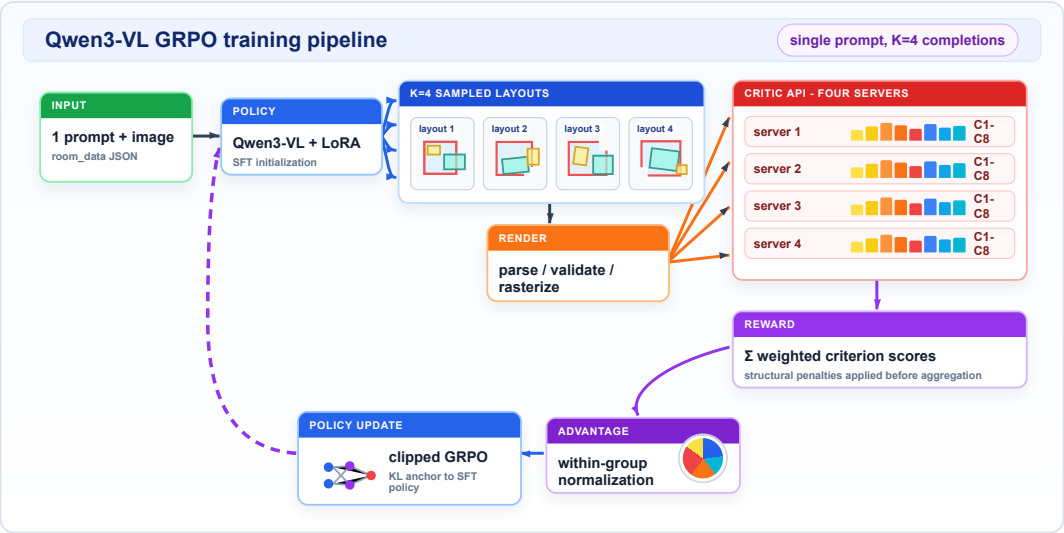


Figure 4: **GRPO pipeline,** $K{=}4$**.** The generator produces four candidate layouts per prompt; each is rendered and scored by distributed critic instances. Within-group advantages drive the policy update.

4.1 Stage 1: SFT Generator Quality

On the held-out 8,001-room test set we report three geometric metrics: out-of-boundary rate (OOB) and furniture intersection rate (FI), both minimised, and ground-truth IoU (GT IoU), maximised. Table 2 summarises six LoRA-tuned models against six zero-shot baselines on the same backbones. Qwen3-VL-8B leads on OOB (2.05 %) and FI (2.46 %); Qwen2.5-14B leads on GT IoU (0.611). Every LoRA model beats its zero-shot counterpart on at least two of the three metrics, confirming the value of polygon-vertex supervision for reducing geometric violations.

Table 3 places our best Stage 1 model against prior layout methods. The baselines (Çelen et al., 2025; Feng et al., 2023; Tang et al., 2024; Wei et al., 2023) emit bounding boxes under different evaluation protocols, so the table is for reference only. FURLA pairs strong OOB/FI (2.05/2.46 %) with competitive GT IoU (0.510); LayoutGPT and I-Design exhibit high OOB or FI on our set, and the diffusion baselines trade lower violations for lower GT IoU, consistent with their different output format.

4.2 Stage 2: Critic Fine-tuning

We evaluated 36 VLM configurations across Qwen3-VL (Bai et al., 2025a), Qwen2.5-VL (Bai et al., 2025b), and InternVL3.5 (Wang et al., 2025a) (3 B–32 B) using two complementary evaluation strategies.

Per-criterion evaluation. Each of the eight criteria is scored in a dedicated VLM call (8 calls/layout), producing criterion-specific feedback and maximising agreement with experts. We tested 18 configurations (9 LoRA, 9 zero-shot). LoRA delivers +57 % QWK (0.421→0.663, fair → substantial agreement), −48 % MAE, and +152 % METEOR. All 9 LoRA models reach QWK > 0.61. Cohen's $d{=}2.41$ (Appendix G.2) is a very large effect. The full 18-model breakdown is in Table 4; the best setup (Qwen3-VL-32B, LoRA $r{=}16$) is used as the Stage 3 critic.

Each per-criterion call returns a 1–4 score plus a short *good/bad/how-to-fix* comment; this actionable score–comment output is the reward signal used in Stage 3 (a worked example is in Appendix 9).

Aggregated evaluation. A single VLM call scores all eight criteria simultaneously (1 call/layout). The aggregated strategy yields only +9 % QWK (0.420→0.458); all models remain in the moderate-agreement tier, none reaching substantial. The full 18-model table is in Appendix F.

Accuracy–efficiency trade-off. The 8× call overhead of per-criterion evaluation buys a +45 % relative QWK gain over the best aggregated model (0.458→0.663) — a Landis–Koch tier change. Table 5 quantifies the degradation: −31 % QWK, −34 % Spearman, +54 % MAE, at 8× cheaper inference. Per-criterion is preferred for quality-critical use; aggregated is viable for high-throughput screening. Error patterns and effect

Table 2: **Stage 1 SFT results.** LoRA fine-tuning vs. zero-shot for the same six backbones. Best per metric in **bold**. Across both groups LoRA reduces OOB from 7.94 % to 2.05 % and FI from 5.43 % to 2.46 % on the best backbone (Qwen3-VL-8B).

Type	Model	OOB (%)↓	FI (%)↓	GT IoU↑
LoRA fine-tuned (6 configurations)				
LoRA	**Qwen3-VL-8B** (r=32)	**2.05**	**2.46**	0.510
LoRA	Qwen3-VL-32B (r=32)	3.05	3.93	0.202
LoRA	Qwen3-VL-4B (r=32)	15.42	7.98	0.254
LoRA	Qwen2.5-14B (r=32)	4.88	6.01	**0.611**
LoRA	Qwen3-30B (r=32)	3.74	5.84	0.090
LoRA	Qwen3-4B (r=32)	4.46	5.47	0.390
Zero-shot baseline (6 configurations)				
Baseline	Qwen3-VL-8B	24.10	14.09	0.070
Baseline	Qwen3-VL-32B	**7.94**	**5.43**	0.110
Baseline	Qwen3-VL-4B	32.82	19.11	0.020
Baseline	Qwen2.5-14B	14.02	9.88	**0.290**
Baseline	Qwen3-30B	10.72	6.85	0.050
Baseline	Qwen3-4B	37.48	10.35	0.060

Table 3: **Comparison with prior layout methods.** FURLA emits full polygon vertices; the baselines emit axis-aligned bounding boxes. Best per metric in **bold**.

Method	OOB↓	FI↓	GT IoU↑
FURLA (Ours)	**2.05**	**2.46**	**0.510**
I-Design (Çelen et al., 2025)	4.51	22.30	0.450
LayoutGPT (Feng et al., 2023)	78.10	12.74	0.080
DiffuScene (Tang et al., 2024)	0.18	0.11	0.410
LEGO-Net (Wei et al., 2023)	0.17	0.05	0.360

sizes are in Appendix G.

4.3 Stage 3: GRPO Reward Optimization

The GRPO run lasted 32 hours on 8×H100-80 GB over 800 training prompts (K=4). Figure 5 shows validation reward rising from ≈0.68 to 0.72 before stabilising. Table 6 reports per-criterion normalised scores (Equation (4)) before and after GRPO. All eight criteria improve; the largest gains are on C7 Natural Light (+2.4 p.p.), C5 Furniture Pairs (+2.2 p.p.) and C2 Focus Points (+1.5 p.p.) — the policy learns to favour daylight, group paired items, and orient seating toward focal elements. Aggregate reward rises from 0.946 to 0.959 (+1.3 p.p.), confirming the critic as a useful learning signal beyond the SFT prior.

Qualitative Stage 3 generation examples are shown in Figure 11.

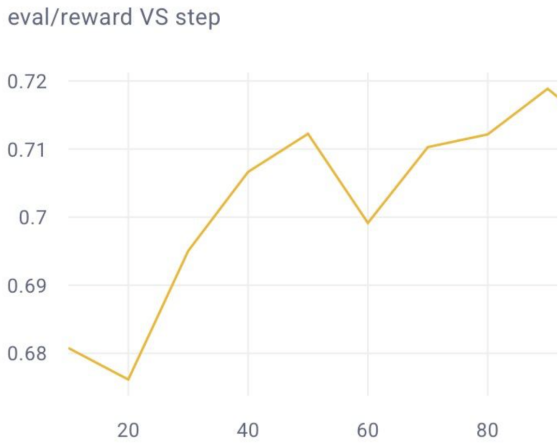


Figure 5: **Validation reward vs. training step.** Reward increases from ≈0.68 to 0.72 and stabilises, confirming progressive policy improvement.

5 Discussion

5.1 Interpretation of Results

The three-stage pipeline produces a coherent picture of human-aligned furniture layout generation in irregular rooms. Stage 1 SFT dramatically reduces geometric violations relative to zero-shot baselines, confirming that polygon-vertex supervision enables models to internalise boundary and clearance constraints that bounding-box methods structurally cannot enforce. Stage 2 establishes that per-criterion LoRA fine-tuning yields a critic whose agreement with expert annotations exceeds the human inter-annotator ceiling; the aggregated strategy trades this accuracy for lower computational cost, suitable for high-throughput screening. Stage 3 demonstrates that the eight-criterion critic provides an informative RL signal: the policy improves on all criteria under GRPO, with the largest gains on natural light, furniture grouping, and walkways—dimensions where the SFT prior was weakest. The error-pattern analysis (overprediction bias, adjacent-or-exact parity with humans)

Table 4: **Per-criterion evaluation (18 models, 4 metrics).** Each criterion is evaluated in a dedicated call (8 calls/layout). LoRA fine-tuning yields $+57\%$ QWK, -48% MAE, and $+152\%$ METEOR. All LoRA models reach substantial agreement (QWK > 0.61).

Type	Model	B	MAE↓	QWK↑	Spearman↑	METEOR↑
LoRA fine-tuned (9 configurations)						
LoRA	**Qwen3-VL-32B** (r=16)	32	**0.3696**	**0.6629**	**0.7554**	0.5739
LoRA	Qwen2.5-VL-32B (r=16)	32	0.3816	0.6587	0.7469	0.5703
LoRA	Qwen3-VL-8B (r=64)	8	0.3789	0.6490	0.7406	0.5744
LoRA	Qwen3-VL-8B (r=16)	8	0.3805	0.6460	0.7415	**0.5840**
LoRA	Qwen3-VL-4B (r=16)	4	0.3809	0.6399	0.7386	0.5779
LoRA	InternVL3.5-8B (r=16)	8	0.3919	0.6346	0.7173	0.5550
LoRA	Qwen2.5-VL-7B (r=16)	7	0.3981	0.6244	0.7222	0.5702
LoRA	Qwen2.5-VL-3B (r=16)	3	0.3860	0.6237	0.7417	0.5760
LoRA	InternVL3.5-4B (r=16)	4	0.3876	0.6122	0.7361	0.5437
Zero-shot baseline (9 configurations)						
Baseline	Qwen3-VL-4B	4	**0.7067**	**0.4210**	0.4665	0.2277
Baseline	Qwen3-VL-32B	32	0.7329	0.4169	**0.4903**	0.2537
Baseline	Qwen2.5-VL-32B	32	0.9032	0.3323	0.4467	**0.2624**
Baseline	InternVL3.5-8B	8	1.0422	0.2462	0.4311	0.2012
Baseline	Qwen3-VL-8B (run 1)	8	0.9603	0.2313	0.3377	0.2278
Baseline	Qwen3-VL-8B (run 2)	8	0.9612	0.2293	0.3358	0.2280
Baseline	InternVL3.5-4B	4	1.0222	0.2116	0.3372	0.2515
Baseline	Qwen2.5-VL-7B	7	1.0410	0.1429	0.2733	0.2238
Baseline	Qwen2.5-VL-3B	3	1.2469	0.1306	0.2586	0.2210

Table 5: **Critic accuracy–efficiency trade-off (Stage 2).** Per-criterion (PC) makes 8 calls per layout; aggregated (Agg.) makes 1. The Δ row reports Agg. relative to PC. The $+45\%$ QWK gain crosses the moderate-to-substantial Landis–Koch threshold.

Approach	MAE↓	QWK↑	ρ↑	Cost
Per-criterion	**0.370**	**0.663**	**0.755**	8×
Aggregated	0.571	0.458	0.495	1×
Δ (Agg. – PC)	$+54\%$	-31%	-34%	8× faster

Table 6: **Per-criterion normalised scores before and after GRPO (Stage 3, held-out test set).** Scores are in $[0, 1]$ via Equation (4). GRPO improves all eight criteria; largest gains on C7, C5, C2.

Criterion	Before	After	Δ
C1 Zones	0.950	0.962	0.012
C2 Focus	0.945	0.960	0.015
C3 Walkways	0.981	0.993	0.012
C4 Access	0.962	0.974	0.012
C5 Pairs	0.916	0.938	0.022
C6 Balance	0.880	0.881	0.001
C7 Light	0.963	0.987	**+0.024**
C8 Entry	0.974	0.980	0.006
Aggregate reward	**0.946**	**0.960**	**+0.014**

indicates that the critic's behaviour is structured and calibratable rather than noisy, supporting its use as an automated reward in production workflows.

A distinguishing practical advantage of the FURLA formulation is its treatment of doors and windows as fixed, pre-positioned inputs. In real residential construction, opening locations are determined before interior design begins: they depend on structural load paths, building-code fire-egress requirements, insolation calculations, and ventilation norms—parameters that furniture arrangement must respect but cannot alter. By conditioning on these architecturally valid openings rather than generating layouts in abstract rectangular rooms, FURLA bridges the gap between academic layout generation and the constraints faced by practising architects and interior designers.

Future work. The main next steps are a learned geometric-validity critic, larger GRPO runs, 3D/BIM export, and regulatory-compliance scoring; details are in Appendix J.

6 Related Work

FURLA sits at the intersection of indoor scene synthesis, VLM-based layout generation, and learned evaluation. Earlier generators such as ATISS (Paschalidou et al., 2021), LayoutGPT (Feng et al., 2023), LLplace (Yang et al.,

2024), OptiScene (Yang et al., 2025), DirectLayout (Ran et al., 2025), and ReSpace (Bucher and Armeni, 2025) mainly assume rectangular or convex rooms, box-based objects, or offline preference labels. Evaluation work such as SceneEval (Tam et al., 2025) and LayoutVLM (Sun et al., 2025) checks physical plausibility but does not train a human-aligned multi-criterion reward model. In contrast, FURLA uses full polygon vertices in irregular rooms, validates an eight-criterion ergonomic rubric with expert agreement, and closes the loop with online GRPO over $K=4$ sampled layouts. Appendix H gives the full comparison.

7 Conclusion

We presented FURLA, a three-stage pipeline for furniture layout generation in real-world irregular rooms with human-aligned ergonomic rewards. Polygon-vertex output paired with SFT supervision reduces out-of-boundary and furniture-intersection rates by 74 % and 55 % over zero-shot baselines, enabling exact wall-contact and door-clearance reasoning that bounding-box representations cannot provide. A LoRA-tuned VLM critic trained on an eight-criterion expert-annotated rubric achieves QWK = 0.663—substantial agreement that matches the human inter-annotator ceiling—and serves as a reliable automated reward signal. An online GRPO loop closing the generator–critic loop yields a further 1.4 % gain in mean ergonomic reward across all eight criteria with no additional human annotation. Taken together, FURLA closes the gap between raw model output and human-aligned interior design in the irregular, architecturally realistic rooms that define real residential spaces.

Limitations

The critic is trained to evaluate ergonomic quality—functional zones, focal points, walkways, accessibility, furniture grouping, spatial balance, natural light, and entry visibility—but is not trained to detect or penalise furniture intersections and out-of-boundary placements; those are handled by explicit structural penalties in the GRPO loop. Generation quality also degrades when rooms contain more than roughly ten furniture items, where maintaining both valid placements and coherent activity zones becomes substantially harder. The critic has an 8:1 overprediction bias (Table 16), which can inflate rewards for mediocre layouts and should be calibrated before deployment. The current GRPO experiment uses 800 training prompts and one 32-hour run, so the observed reward plateau at 0.72 should be read as a resource-limited pilot rather than a convergence ceiling. Finally, FURLA operates on top-down 2D floor plans and does not yet model 3D constraints such as furniture height, sightlines, or natural light penetration angles; the full pipeline also requires substantial GPU resources (8×H100-80 GB), which limits reproducibility.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, and 1 others. 2025a. Qwen3-VL technical report. *arXiv preprint*. ArXiv:2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025b. Qwen2.5-VL technical report. *arXiv preprint*. ArXiv:2502.13923.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Martin J. J. Bucher and Iro Armeni. 2025. ReSpace: Text-driven 3d indoor scene synthesis and editing with preference alignment. *arXiv preprint*. ArXiv:2506.02459.

Ata Çelen, Guo Han, Konrad Schindler, Luc Van Gool, Iro Armeni, Anton Obukhov, and Xi Wang. 2025. I-Design: Personalized llm interior designer. In *Computer Vision – ECCV 2024 Workshops*, volume 15624 of *LNCS*, pages 217–234. Springer.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.*, 70(4):213–220.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition. Lawrence Erlbaum.

Tri Dao. 2023. FlashAttention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint*. ArXiv:2307.08691; `https://github.com/Dao-AILab/flash-attention`.

Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2023. LayoutGPT: Compositional visual planning and generation with large language models. In *NeurIPS*. ArXiv:2305.15393.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *ICLR*. ArXiv:2106.09685.

Xin Hu, Pengfei Xu, Jin Zhou, Hongbo Fu, and Hui Huang. 2025. StructLayoutFormer: Conditional structured layout generation via structure serialization and disentanglement. *IEEE Trans. Vis. Comput. Graph.* ArXiv:2510.26141.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Despoina Paschalidou, Amlan Kar, Maria Shugrina, Karsten Kreis, Andreas Geiger, and Sanja Fidler. 2021. ATISS: Autoregressive transformers for indoor scene synthesis. In *NeurIPS*. ArXiv:2110.03675.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, volume 36, pages 53728–53741. ArXiv:2305.18290.

Xingjian Ran, Yixuan Li, Linning Xu, Mulin Yu, and Bo Dai. 2025. DirectLayout: Direct numerical layout generation for 3d indoor scene synthesis via spatial reasoning. *arXiv preprint*. ArXiv:2506.05341.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proc. 26th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, pages 3505–3506. ACM. https://github.com/microsoft/DeepSpeed.

Mohammad Amin Shabani, Sepidehsadat Hosseini, and Yasutaka Furukawa. 2023. HouseDiffusion: Vector floorplan generation via a diffusion model with discrete and continuous denoising. In *CVPR*, pages 5466–5475.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junjie Song, Xinchi Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and 1 others. 2024. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint*. ArXiv:2402.03300.

Fan-Yun Sun, Weiyu Liu, Siyi Gu, Dylan Lim, Goutam Bhat, Federico Tombari, Manling Li, Nick Haber, and Jiajun Wu. 2025. LayoutVLM: Differentiable optimization of 3d layout via vision-language models. In *CVPR*. ArXiv:2412.02193.

Hou In Ivan Tam, Hou In Derek Pun, Austin T. Wang, Angel X. Chang, and Manolis Savva. 2025. SceneEval: Evaluating semantic coherence in text-conditioned 3d indoor scene synthesis. *arXiv preprint*. ArXiv:2503.14756.

Jiapeng Tang, Yinyu Nie, Lev Markhasin, Angela Dai, Justus Thies, and Matthias Nießner. 2024. DiffuScene: Denoising diffusion models for generative indoor scene synthesis. In *CVPR*, pages 20507–20518.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouedec. 2020. TRL: Transformer reinforcement learning. https://github.com/huggingface/trl.

Can Wang, Hongliang Zhong, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. 2024. Chat2Layout: Interactive 3d furniture layout with a multimodal llm. *arXiv preprint*. ArXiv:2407.21333.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhe Chen, and 1 others. 2025a. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint*. ArXiv:2508.18265; https://arxiv.org/abs/2508.18265.

Xiaoyan Wang, Zeju Li, Yifan Xu, Jiaxing Qi, Zhifei Yang, Ruifei Ma, Xiangde Liu, and Chao Zhang. 2025b. Spatial 3D-LLM: Exploring spatial awareness in 3d vision-language models. *arXiv preprint*. Accepted ICME 2025; arXiv:2507.16524.

Qiuhong Anna Wei, Sijie Ding, Jeong Joon Park, Rahul Sajnani, Adrien Poulenard, Srinath Sridhar, and Leonidas Guibas. 2023. LEGO-Net: Learning regular rearrangements of objects in rooms. In *CVPR*, pages 19037–19047.

Yixuan Yang, Junru Lu, Zixiang Zhao, Zhen Luo, James J. Q. Yu, Victor Sanchez, and Feng Zheng. 2024. LLplace: The 3d indoor scene layout generation and editing via large language model. *arXiv preprint*. ArXiv:2406.03866.

Yixuan Yang, Zhen Luo, Tongsheng Ding, Junru Lu, Mingqi Gao, Jinyu Yang, Victor Sanchez, and Feng Zheng. 2025. OptiScene: Llm-driven indoor scene layout generation via scaled human-aligned data synthesis and multi-stage preference optimization. *arXiv preprint*. ArXiv:2506.07570.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proc. 62nd Annual Meeting of the Assoc. for Computational Linguistics (System Demonstrations)*, pages 400–410. Association for Computational Linguistics. ArXiv:2403.13372.

Appendix

The appendix is ordered to mirror the main paper: data and metrics first, then training setup (hyperparameters and prompts), then the GRPO algorithm, and finally extended results and analysis.

- Sec. A– Extended Dataset and Training Details (splits, routing)
- Sec. B – Evaluation Metric Definitions (Stages 1–3)
- Sec. C – Hyperparameters and Training Configurations (all stages)
- Sec. D – Prompt Templates (generator, critic, GRPO)
- Sec. E – GRPO Per-Step Algorithm
- Sec. F – Extended Results (full aggregated critic table)
- Sec. G – Extended Critic Analysis (effect sizes, errors, scaling)
- Sec. H – Qualitative Examples and Extended Related Work
- Sec. I – Licenses and Intended Use
- Sec. J – Future Work and Planned Improvements

A Extended Dataset and Training Details

This section supports the main-text description of the three stages by giving (i) dataset scale and composition, (ii) record structure and vocabulary, (iii) split statistics and exact counts, and (iv) distributed critic routing for Stage 3. We first show the Stage 1 dataset overview and room-type coverage, then the record format and splits, and finally the routing table.

Dataset scale and room types. Figure 6 presents three complementary views of the Stage 1 corpus. The left panel shows the total counts (53,332 rooms drawn from 22,000 real apartments), the middle panel shows the per-type room-count distribution—bedroom and living room dominate, while technical rooms, terraces, and wardrobe rooms are rare tail classes—and the right panel illustrates the distribution of furniture counts per room, peaking around 4–7 items. Together these panels explain the class-imbalance challenge and motivate the 70/15/15 stratified split described in Table 7.

Figure 7 shows one fully annotated example for each of the 13 room types. Each image renders the raw floor-plan polygon (grey walls), door/window boxes (blue rectangles), and the ground-truth placed furniture polygons (coloured outlines). The diversity is substantial: bedrooms are roughly rectangular with 4–6 items; open-concept rooms are large irregular polygons with 10+ items; terraces and corridors are narrow elongated spaces with 1–3 items. This heterogeneity is precisely why a model must learn to reason about irregular boundaries rather than relying on rectangular-room heuristics.

A.1 Dataset Record Structure

Each Stage 1 room record contains: ordered (x, y) room-boundary polygon vertices; door and window bounding boxes; canonical (unplaced) furniture polygon templates with the longest side horizontal and bottom-left corner at $(0, 0)$; and a ground-truth placed layout consisting of vertex coordinates for every furniture item after rotation and translation, serialized as JSON.

The vocabulary of furniture items covers 45 categories: armchair, bar counter, bar stool, bathtub, bookcase, cabinet, chaise lounge, chair, coffee table, desk, dining table, dishwasher, double bed, dresser, floor lamp, floor plant, floor vase, home sauna, kitchen cabinet, kitchen island, lamp, mirror, nightstand, office chair, ottoman, oven, plant, rug, refrigerator, shelving unit, shoe cabinet, shower, sink, single bed, sofa, stove, table lamp, toilet, TV, TV stand, wardrobe, washing machine, and coat rack. The 13 room types are: balcony, bathroom, bedroom, children's room, corridor, hallway, kitchen, living room, office, open-concept living room, technical room, terrace, and wardrobe room.

A.2 Dataset Splits

Figure 8 verifies that the three splits are statistically indistinguishable across six geometry dimensions, ruling out any systematic data-leakage or selection bias. The top row shows *composition* statistics: the average number of furniture items per room ($\approx$6.1, all splits), the mean room footprint area ($\approx$8.7 m^2), and the fraction of the room area covered by furniture ($\approx$40.6 %). The bottom row shows *quality* statistics: the mean number of polygon vertices per room boundary (a proxy for irregular shape complexity, $\approx$6.7), the fraction of rooms whose ground-truth annotation already has out-of-boundary furniture ($\approx$5 %), and the fraction with furniture overlaps ($\approx$3 %). The near-identical

bar heights across Train/Val/Test in every panel confirm that stratified partitioning preserved the joint distribution of room complexity, size, and annotation artifacts.

Table 7 gives the exact room counts for the Stage 1 SFT generator training. A 70/15/15 split on 53,332 rooms yields 37,330 training rooms and two 8,001-room held-out sets; the validation set is used for early stopping and the test set is held out until final evaluation.

Table 7: **Stage 1 dataset splits** for SFT layout generation (70 / 15 / 15 partition).

Split	Rooms	Fraction
Train	37,330	70 %
Validation	8,001	15 %
Test	8,001	15 %
Total	**53,332**	**100 %**

Table 8 details the Stage 2 critic splits for both evaluation strategies. Under the per-criterion strategy, each of the 1,121 annotated layout images produces up to 8 separate criterion records (one per criterion), yielding 20,160 training records in total with an average applicability of 80.5 % (some criteria, such as C5 furniture pairs, are not applicable to rooms with no relevant furniture pairs). Under the aggregated strategy, all applicable criteria are bundled into a single record per image, giving 2,242 records—roughly 9× fewer than per-criterion—which explains why the aggregated backbone achieves similar QWK to the per-criterion baseline with far less data.

Table 8: **Stage 2 dataset splits — both critic strategies.** Both strategies use the same 1,121 annotated images (754/183/184 train/val/test). Per-criterion: one training record per criterion per image (8 records per image, filtered by applicability). Aggregated: one training record per image with all applicable criteria bundled together.

	Per-criterion		**Aggregated**	
Split	Records	Appl. %	Records	Images
Train	14,112	79.3	1,508	754
Validation	3,024	85.2	366	183
Test	3,024	81.2	368	184
Total	**20,160**	**80.5**	**2,242**	**1,121**

Table 9 describes the Stage 3 GRPO prompt set. Only 1,000 distinct room prompts are used for GRPO (800 train and 200 validation), because each prompt generates $K{=}4$ layout candidates per training step, giving 3,200 candidate evaluations per epoch in training alone. The smaller size compared to Stage 1 is intentional: GRPO fine-tuning requires the critic to score every generated candidate at every step, so the bottleneck is inference throughput rather than data diversity; 800 unique rooms provide sufficient variance for meaningful advantage estimation across the 32-hour training run.

Table 9: **Stage 3 GRPO prompt dataset.** Each prompt generates $K{=}4$ candidate layouts for reward computation.

Split	Prompts	Candidates ($K{=}4$)
Train	800	3,200
Validation	200	800
Total	**1,000**	**4,000**

B Evaluation Metric Definitions

The main text (§4) reports Stage 1 geometric metrics (OOB, FI, GT IoU), Stage 2 critic metrics (MAE, QWK, Spearman, METEOR, Cohen's d), and Stage 3 reward. Here we give the formal definitions used in the paper.

B.1 Stage 1 Geometric Metrics

Let $\mathcal{L}{=}\{P_l\}_{l=1}^{M}$ be a predicted layout of M furniture polygons and $\mathbf{P}^{\text{room}}$ the room boundary.

Out-of-Boundary Rate (OOB). The fraction of furniture items whose placed polygon is not fully contained within the room boundary:

$$\text{OOB}(\mathcal{L}) = \frac{1}{M}\sum_{l=1}^{M} \mathbf{1}\big[P_l \not\subseteq \mathbf{P}^{\text{room}}\big]. \quad (5)$$

Furniture Intersection Rate (FI). The fraction of furniture pairs that overlap (non-exempt items only):

$$\text{FI}(\mathcal{L}) = \frac{2}{M(M-1)}\sum_{i<j} \mathbf{1}\big[\text{area}(P_i \cap P_j) > 0\big]. \quad (6)$$

Ground-Truth IoU (GT IoU). Given ground-truth layout $\mathcal{L}^*{=}\{P_l^*\}$, the mean intersection-over-union between predicted and reference polygons, matched by object identity:

$$\text{GT IoU}(\mathcal{L}, \mathcal{L}^*) = \frac{1}{M}\sum_{l=1}^{M} \frac{\text{area}(P_l \cap P_l^*)}{\text{area}(P_l \cup P_l^*)}. \quad (7)$$

B.2 Stage 2 Critic Metrics

Let $\hat{s}_c \in \{1, \ldots, K_c\}$ be the model's predicted score for criterion c and s_c^* the expert label.

Mean Absolute Error (MAE).

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |\hat{s}_{c,i} - s^*_{c,i}|, \tag{8}$$

where N is the number of evaluated criterion–image pairs.

Quadratic Weighted Kappa (QWK). QWK (Cohen, 1968) measures inter-rater agreement with quadratic penalization of ordinal disagreements. Let O be the observed $K_c \times K_c$ confusion matrix and E the expected matrix under marginal independence:

$$\mathrm{QWK} = 1 - \frac{\sum_{i,j} w_{ij} O_{ij}}{\sum_{i,j} w_{ij} E_{ij}}, \tag{9}$$

$$w_{ij} = \frac{(i-j)^2}{(K_c - 1)^2}. \tag{10}$$

Landis–Koch interpretation (Landis and Koch, 1977): $<$ 0.0 poor, 0.0–0.20 slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, $>$ 0.80 almost perfect.

Spearman Rank Correlation (ρ).

$$\rho = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)}, \tag{11}$$

where d_i is the difference in ranks between $\hat{s}_{c,i}$ and $s^*_{c,i}$.

METEOR. METEOR (Banerjee and Lavie, 2005) evaluates the textual similarity of model-generated critiques to expert comments. For hypothesis h and reference r:

$$\mathrm{METEOR}(h, r) = F_{\mathrm{mean}} \cdot (1 - \mathrm{Pen}), \tag{12}$$

$$F_{\mathrm{mean}} = \frac{10\,P\,R}{R + 9P}, \tag{13}$$

where P and R are unigram precision and recall over aligned token pairs, and $\mathrm{Pen}{=}0.5 \cdot (\mathrm{chunks}/\mathrm{matches})^3$ penalizes fragmented alignments. We compute METEOR per criterion and report the macro-average across all eight criteria.

Cohen's d (effect size).

$$d = \frac{\mu_1 - \mu_2}{s_p}, \tag{14}$$

$$s_p = \sqrt{\frac{(n_1 - 1)\,s_1^2 + (n_2 - 1)\,s_2^2}{n_1 + n_2 - 2}}, \tag{15}$$

where μ_i, s_i, n_i are the mean, standard deviation, and sample size of each group (LoRA vs. baseline; per-criterion vs. aggregated) (Cohen, 1988). Conventional thresholds: $d{=}0.2$ small, 0.5 medium, 0.8 large, 2.0 very large.

C Hyperparameters and Training Configurations

This section gives the full hyperparameter and infrastructure settings for each stage, referenced from §3.

C.1 Stage 1: SFT Generator

Table 10 lists the complete hyperparameter configuration for all Stage 1 SFT experiments. All models are trained with LLaMA-Factory (Zheng et al., 2024) using DeepSpeed ZeRO-3 and bfloat16 precision. LoRA adapters target all attention projection (Q, K, V, O) and MLP modules (`lora_target: all`). No weight decay is applied; AdamW epsilon uses the PyTorch default (10^{-8}). Several design choices in the table deserve explanation. We fix LoRA rank $r{=}32$ for all models, balancing expressivity with GPU memory budget; a larger $r{=}64$ increases task-specific parameters by $4\times$ but yields at most $+0.3$ % QWK in our ablations, so $r{=}32$ is production-optimal. Setting $\alpha{=}2r{=}64$ keeps the effective scaling factor at 2.0 regardless of rank, a common practice that avoids retuning learning rate when rank changes. Fifteen epochs with 10 % warm-up and cosine decay proved sufficient for all backbones to plateau (monitored on the validation set); further epochs do not improve GT IoU and begin to hurt OOB. VL models receive floor-plan images capped at 262,144 pixels (512×512 effective resolution), which captures furniture polygon detail while keeping sequence lengths manageable.

C.2 Stage 2: VLM Critic

Table 11 lists the complete hyperparameter configuration for all Stage 2 critic experiments. The critic backbone is always Qwen3-VL-32B for the best-performing model; the full ablation covers Qwen3-VL, Qwen2.5-VL, and InternVL3.5 families (3 B–32 B). All experiments use the same optimizer, schedule, and inference parameters; only LoRA rank varies in the rank-ablation runs. Two entries in the table require clarification. The *per-criterion* strategy uses LoRA $r{=}16$, whereas *aggregated* uses $r{=}64$: aggregated must compress eight criteria into one response, so larger capacity

Table 10: **Stage 1 SFT hyperparameters.** All models use AdamW, cosine LR schedule, LoRA (α=2r), DeepSpeed ZeRO-3 (Rasley et al., 2020), bfloat16, FlashAttention-2 (Dao, 2023). VL models additionally use `image_max_pixels` = 262,144.

Model	Type	r	α	LR	Epochs	Eff. batch	Warmup
Qwen3-VL-8B	VL (ours best)	32	64	2×10^{-4}	15	4	10%
Qwen3-VL-32B	VL	32	64	2×10^{-4}	15	8	10%
Qwen3-VL-4B	VL	32	64	2×10^{-4}	15	4	10%
Qwen2.5-14B	LLM	32	64	2×10^{-4}	15	4	10%
Qwen3-30B	LLM	32	64	2×10^{-4}	15	8	10%
Qwen3-4B	LLM	32	64	2×10^{-4}	15	8	10%

is needed to match quality; at r=16 the aggregated QWK drops by -0.04 absolute. Inference temperature is set low (0.3) to reduce stochastic variance in the ordinal score, while a mild repetition penalty (1.01) suppresses score-label repetition artifacts without distorting the distribution. Per-criterion calls are capped at 1,024 tokens (score + 2–3 sentence comment), whereas aggregated calls need 3,200 tokens to fit all eight score–comment pairs without truncation.

C.3 Stage 3: GRPO

Table 12 lists the complete configuration for the Stage 3 GRPO training run. Several GRPO-specific choices are worth highlighting. The policy learning rate (5×10^{-7}) is set two orders of magnitude lower than Stage 1 SFT, because the policy is already near-optimal and large steps cause catastrophic forgetting of the structural constraints learned at Stage 1. The KL penalty β=0.01 keeps the policy close to the SFT reference, preventing reward hacking (e.g., generating empty layouts that trivially satisfy geometric constraints). Reward sharpness γ=1.5 amplifies the advantage signal within each group of K=4 candidates: without sharpening, per-group reward variance is small ($\approx$0.02 std) and gradient magnitude is negligible; with γ=1.5, effective variance grows to $\approx$0.05, matching the scale needed for stable GRPO updates. The violation threshold of > 3 total geometric violations skips the expensive critic call for badly malformed layouts, assigning a minimal reward of 0.05 and saving $\approx$30 % of critic API calls without meaningfully changing the reward distribution. Per-device batch size 1 with 16 gradient accumulation steps compensates for the fixed per-GPU memory budget on H100-80 GB, where a single multimodal GRPO forward pass (prompt + 4 completions) already occupies 50–60 GB.

C.4 Distributed Critic Routing

For Stage 3, each GRPO step must score K=4 candidate layouts on all eight criteria, meaning 32 critic calls per prompt. Serialising these calls on a single GPU would be the dominant latency bottleneck. We therefore distribute the eight criteria across four dedicated critic API servers in a round-robin fashion, assigning two criteria per server. Table 13 documents the mapping: server 8000 (GPU 0) handles C1 (Functional Zones) and C5 (Furniture Pairs); server 8001 handles C2 and C6; server 8002 handles C3 and C7; server 8003 handles C4 and C8. With four servers running in parallel, the effective per-step critic latency equals the time for one server to score K=4 images on two criteria, cutting scoring time by roughly 4$\times$ compared to a single-server baseline. Grouping complementary criteria per server (e.g., spatial structure C1 with pair-placement C5) ensures each server's outputs remain semantically coherent, simplifying reward aggregation.

D Prompt Templates

The following templates are referenced in §3 (Stage 1 and Stage 2) and in the GRPO algorithm (§E). We show one example per stage.

D.1 Stage 1: Layout Generator Prompt

All Stage 1 SFT training and inference use a fixed system prompt combined with a per-room user message containing the floor-plan image and a JSON room payload. The system prompt encodes four classes of knowledge that the model must learn: (1) *geometric constraints* (rules 6, 7—all polygons must be inside the room; non-exempt furniture must not overlap); (2) *ergonomic heuristics* (rules 1–5, 8–10—place items along walls, chairs face tables, leave walking space, avoid blocking doors); (3) *coordinate convention* (x increases rightward, y increases upward, initial polygon with longest side

Table 11: **Stage 2 critic hyperparameters.** Per-criterion and aggregated strategies share the same training setup; inference parameters differ.

Parameter	Value
Training	
Optimizer	AdamW
Learning rate	2×10^{-4}
LR schedule	Cosine
Warmup ratio	10%
LoRA rank r (per-criterion best)	16
LoRA rank r (aggregated best)	64
LoRA α	$2r$
LoRA targets	All attention + MLP
Training pairs (per-criterion)	14,112
Training records (aggregated)	1,508
Precision	bfloat16
Inference	
Temperature	0.3
Top-p	0.9
Repetition penalty	1.01
Max tokens (per-criterion)	1,024
Max tokens (aggregated)	3,200

horizontal and bottom-left at $(0, 0)$); and (4) *output format* (a JSON list of `{"object", "position": {"x", "y", "angle"}}` entries). Structuring the prompt this way lets the SFT model learn a strong syntactic prior (valid JSON, correct key names) before Stage 3 GRPO introduces semantic refinement pressure from the critic.

Stage 1: Layout Generator — System Prompt

```
You are a skilled room layout designer. Your task is to
arrange [Objects] within a given [Room Type] effectively.
Follow these guidance to complete your design: (1) Extract
the [Room Type], [Vertices of Room Polygon], [Objects], and
their [Vertices of Object Polygon] from the provided JSON
data. (2) Analyze room shape and positions of doors and
windows from the provided image. (3) Analyze the spatial
relationships among [Objects] within the specified [Room
Type]. (4) Determine and design the precise location of all
[Objects] inside [Vertices of Room Polygon]. Check object
polygons do not overlap and that the layout is functional
and visually appealing. (5) The initial furniture polygons
have longest side horizontal and left bottom corner at (0,
0). When you position polygons, you can rotate and translate
them. (6) Make sure all [Objects] lie fully inside the room.
(7) The order of polygon points is (x, y). (8) You are also
given vertices of polygons for doors, windows and rooms. Do
not occlude them, and leave enough space to open doors.

Rules: 1. Consider spatial factors like accessibility and
aesthetics. 2. Place objects in the corners of the room if
possible to make the room look more spacious. 3. Chairs must
be placed near to the table/desk and face to the table/desk.
4. Leave enough walking space. 5. Place objects along
the walls when possible. 6. All furniture polygons should
be entirely inside room contour. 7. Positioned furniture
polygons should not overlap, except for polygons of the
following categories, which can overlap with any object:
lamp, plant, rug, oven, sink, TV, mirror. 8. Do not occlude
doors and windows, leave enough space in front of them. 9.
Leave enough space in front of furniture with doors/chests
(i.e., cabinets, refrigerators, oven, shelving units). 10.
Do not allow floating objects, like fridge in the middle of
a room. 11. Remove any unnecessary gaps between furniture
items.

In coordinate system of the room, x increases rightwards, y
increases upward.

Report your design with new positions (as list of (x, y)
points) of polygon vertices for each furniture (exclude
windows and doors) in JSON format, as follows: [{"object":
"object", "position": [[x1, y1], [x2, y2], [x3, y3], [x4,
y4]]}, {"object": "object 2", "position": [[x1, y1], [x2,
y2], [x3, y3], [x4, y4]]}]
```

User prompt (image + room payload).

Stage 1: Layout Generator — User Message (example balcony)

```
<image> {'room_type': 'balcony', 'room_vertices':
[[131,650],[1563,650],[1566,194],[129,190]],
'furniture': [{'object': 'armchair', 'points':
[[300,302],[334,34],[29,0],[0,262]]}, {'object': 'coffee
table', 'points': [[247,247],[254,4],[8,0],[0,244]]}],
'doors': [{'object': 'door', 'bbox_points':
[[782,894],[1428,896],[1429,651],[781,650]]}],
'windows': [{'object': 'window', 'bbox_points':
[[131,190],[1566,194],[1565,0],[131,2]]}]}
```

Expected response.

Stage 1: Layout Generator — Model Response

```
[{'object': 'armchair', 'position':
[[155,556],[425,575],[442,269],[179,254]]},
{'object': 'coffee table', 'position':
[[514,532],[757,537],[759,291],[515,285]]}]
```

D.2 Stage 2: Per-Criterion Critic Prompt

Each of the eight criteria is evaluated in a dedicated inference call. The model outputs a JSON object with a numeric score and a short comment. The per-criterion design has two key advantages over aggregating all criteria into one call. First, it forces the model to focus exclusively on one dimension (e.g., C1 functional zones) without distraction from the other seven, reducing the cognitive load and improving calibration accuracy. Second, it produces criterion-specific comments that are action-

Table 12: **Stage 3 GRPO training configuration.** Uses TRL (von Werra et al., 2020) `GRPOTrainer` with the Stage 1 SFT generator (Qwen3-VL-8B, r32) as policy initialization.

Parameter	Value
Optimizer and schedule	
Optimizer	AdamW
Learning rate	5×10^{-7}
LR schedule	Cosine
Warmup ratio	10%
Weight decay	0.01
Batch and sequence	
Per-device train batch size	1
Gradient accumulation steps	16
Effective batch size	16
Candidates per prompt (K)	4
Max prompt length (tokens)	4,096
Max completion length (tokens)	1,024
GRPO-specific	
KL penalty β	0.01
Clip ratio ϵ	0.2 (default)
Reward sharpness γ	1.5
Structural penalty (overlap)	0.05 per item
Structural penalty (boundary)	0.08 per item
Violation threshold (skip critic)	>3 total violations
Infrastructure	
Hardware	8×H100-80 GB
Precision	bfloat16
Attention	FlashAttention-2
Training prompts	800
Total wall time	32 hours
Cycle time per prompt	≈115 s

Table 13: **Critic API round-robin routing (Stage 3).** Eight criteria are distributed across 4 GPU servers; each server evaluates 2 criteria on $K=4$ candidate images per GRPO step.

Server (port / GPU)	Criterion 1	Criterion 2	Codes
8000 / GPU 0	C1 Functional Zones	C5 Furniture Pairs	C1, C5
8001 / GPU 1	C2 Focus Points	C6 Spatial Balance	C2, C6
8002 / GPU 2	C3 Walkways	C7 Natural Light	C3, C7
8003 / GPU 3	C4 Accessibility	C8 Entry Visibility	C4, C8

able feedback ("the desk is not near the window"), whereas aggregated comments tend to be superficial and generic. The instruction explicitly warns the model not to evaluate other criteria ("Evaluate ONLY this criterion"), which reduces score leakage across dimensions observed in early ablations. Each call also returns a 1–4 ordinal score with anchor labels, which aligns with the QWK ordinal metric. Example for criterion C1 (Functional Zones):

Stage 2: Per-Criterion Critic — Prompt (example: C3 Walkways, bathroom)

```
You are a professional interior layout critic.
ORIENTATION: - North (N) is UP, South (S) is DOWN. - Allowed directions: N/S/E/W and diagonal NE/NW/SE/SW.
ROOM TYPE: bathroom
TASK: Evaluate ONLY this criterion: - Criterion ID: c3_walkways - Criterion name: Walkways inside the room
GATING (when not applicable): Always applicable.
DESCRIPTION: Evaluate circulation inside the room: whether key paths between furniture are comfortable and not blocked. All widths must be expressed in step units (e.g., 1/3 step, 2/3 step, 3/4 step).
SCORING SCALE (1-4): 1 — Walkways are blocked (< 1/3 step). 2 — Too narrow (~2/3 step), inconvenient. 3 — Main paths are free with one narrow spot. 4 — All main walkways are > 3/4 step, straight and comfortable.
OUTPUT FORMAT: - Return STRICTLY a JSON object with EXACTLY the keys shown in the schema.
COMMENT STRUCTURE (mandatory): provide 4 sections in "comment": good, bad, how_to_fix, add_or_remove. DIRECTIONS: use ONLY N/S/E/W and diagonals. Forbidden: top/bottom/left/right, here/there. CRITERION INDEPENDENCE: evaluate ONLY the current criterion.
Return your answer STRICTLY as JSON (no extra text): {"criterion_id": "c3_walkways", "applicable": true/false, "score": 1..4 (or null if not applicable), "comment": {"good": "...", "bad": "...", "how_to_fix": "...", "add_or_remove": "..."}, "not_applicable_reason": "..."}
Now look at the image and evaluate this criterion. <image>
```

Expected response.

Stage 2: Per-Criterion Critic — Model Response

{"criterion_id": "c3_walkways", "applicable": true, "score": 4, "comment": {"good": "Wide passages to the furniture.", "bad": "No changes are required.", "how_to_fix": "No changes are required.", "add_or_remove": "No changes are required."}}

D.3 Stage 2: Aggregated Critic Prompt

In the aggregated strategy, all eight criteria are evaluated in a single call. The model receives the layout image and the combined rubric, and returns a JSON object mapping each criterion to a score and comment. This strategy is appealing for its $8\times$ lower inference cost, but it comes with a trade-off: the model must simultaneously attend to all eight dimensions of a single layout image, which increases the output length (up to 3,200 tokens) and requires a larger LoRA rank to maintain calibration quality (see Table 11). The prompt lists the criteria with short inline descriptions and their scoring scales, ordered by the same priority hierarchy as Stage 3 (CRITICAL > HIGH > IMPORTANT), so the model sees which aspects matter most if token budget forces truncation.

Stage 2: Aggregated Critic — Prompt (all 8 criteria in one call)

```
You are a professional interior layout critic.
ORIENTATION: - North (N) is UP, South (S) is DOWN. Use N/S/W/E for directions.
ROOM TYPE: {room_type}
CONTEXT: Evaluate ALL 8 criteria together for this single image in one response. Maintain consistency across criteria: if you mention an object in one criterion, reference the same object in related criteria.
APPLICABILITY RULES: - c1_zones, c2_focus, c4_accessibility, c5_pairs, c6_balance: applicable for living room, open concept living room, kitchen, bedroom, office, balcony, terrace; NOT applicable for bathroom, corridor, hallway, wardrobe room. - c3_walkways: always applicable. - c7_light: only if window exists. - c8_entry: only if door/opening exists.
CRITERIA (all 8): 1. c1_zones [1-4] Functional zones: groups of 2+ items with shared function (lounge, sleep, work...). 2. c2_focus [1-4] Focus points: seating oriented toward focal element (TV, table, desk). TV centred. 3. c3_walkways [1-4] Walkways: passages > 3/4 step comfortable; blocked < 1/3 step. 4. c4_accessibility [1-3] Usability: hinged doors open fully; coffee table reachable. 5. c5_pairs [1-2] Furniture groups: table+chairs, bed+nightstands, sofa+table assembled. 6. c6_balance [1-3] Space balance: 60-70% fill, no dead zones, no overloaded corners. 7. c7_light [1-3] Natural light: key zones near windows; TV not facing windows. 8. c8_entry [1-4] Entry clearance: door opens freely; straight unobstructed path inward.
COMMENT STRUCTURE per criterion: good / bad / how_to_fix / add_or_remove (use "-" if not applicable).
Return STRICTLY as JSON (no extra text): {"room_type": "...", "results": [{"criterion_id": "c1_zones", "applicable": true/false, "score": 1..4, "comment": {"good": "...", "bad": "...", "how_to_fix": "...", "add_or_remove": "..."}}, ...]}
<image>
```

Example response (bathroom — most criteria not applicable).

Stage 2: Aggregated Critic — Model Response

{"room_type": "bathroom", "results": [{"criterion_id": "c1_zones", "applicable": false, "score": null, "comment": {"good":"-","bad":"-","how_to_fix":"-","add_or_remove":"-"}, "not_applicable_reason": "Not applicable: this is bathroom - zoning is not required."}, {"criterion_id": "c3_walkways", "applicable": true, "score": 4, "comment": {"good": "Wide passages to the furniture.", "bad": "No changes are required.", "how_to_fix": "No changes are required.", "add_or_remove": "No changes are required."}}, {"criterion_id": "c8_entry", "applicable": true, "score": 4, "comment": {"good": "Straight unobstructed path from the door.", "bad": "No changes are required.", "how_to_fix": "No changes are required.", "add_or_remove": "No changes are required."}}, ...]}

D.4 Stage 3: GRPO Training Prompt

Each GRPO training example is a self-contained multimodal package that supports (i) sampling K layout candidates and (ii) computing critic-based rewards for advantage estimation. Concretely, every item includes:

- **Prompt** (§D.4): a fixed instruction template that combines (a) layout-generation requirements, (b) the 8-criterion scoring rubric, and (c) strict constraints. The prompt embeds the serialized `room_data` payload and a sample identifier tag (e.g., `<!- METADATA: sample_1318 ->`) for traceability.
- **Room-plan image**: a rasterized floor-plan with walls, doors, and windows, inserted at the `<image>` placeholder and encoded by the vision backbone.
- **`room_data`**: a structured dictionary containing `room_type`, room polygon vertices, canonical (unplaced) furniture polygons, and door/window bounding boxes.
- **`sft_layout`**: the initial placed layout produced by the SFT model; the policy is instructed to *refine* it.
- **Target output**: a JSON list of furniture objects with updated polygon vertices, parsed and validated before reward computation.

$K{=}4$ candidates are sampled per item; invalid generations receive zero reward and are skipped from critic evaluation.

Stage 3: GRPO Policy — Training Prompt (sample_1795, bathroom)

```
You are a skilled room layout designer. Your task is to arrange [Objects] within a given [Room Type] effectively. Follow these guidance to complete your design: (1) Extract the [Room Type], [Vertices of Room Polygon], [Objects], and their [Vertices of Object Polygon] from the provided JSON
```

```
data. (2) <image> Analyze room shape and positions of doors
and windows from the provided image (3) Analyze the spatial
relationships among [Objects] within the specified [Room
Type]. (4) Determine and design the precise location of all
[Objects] inside [Vertices of Room Polygon]. Check object
polygons do not overlap and that the layout is functional
and visually appealing. (5) The initial furniture polygons
have longest side horizontal and left bottom corner at (0,
0). When you position polygons, you can rotate and translate
them. (6) Make sure all [Objects] lie fully inside the room.
(7) The order of polygon points is (x, y). (8) You are also
given vertices of polygons for doors, windows and rooms. Do
not occlude them, and leave enough space to open doors.

Rules: 1. Consider spatial factors like accessibility and
aesthetics. 2. Place objects in the corners of the room if
possible to make the room look more spacious. 3. Chairs must
be placed near to the table/desk and face to the table/desk.
4. Leave enough walking space. 5. Place objects along the
walls when possible. 6. All furniture polygons should be
entirely inside the room contour. 7. Positioned furniture
polygons should not overlap, except for polygons of the
following categories, which can overlap with any object:
lamp, plant, rug, oven, sink, TV, mirror. 8. Do not occlude
doors and windows, leave enough space in front of them. 9.
Leave enough space in front of furniture with doors/chests
(i.e., cabinets, refrigerators, oven, shelving units). 10.
Do not allow floating objects, like fridge in the middle of
a room. 11. Remove any unnecessary gaps between furniture
items.

In coordinate system of the room, x increases rightwards, y
increases upward.

==========
EVALUATION CRITERIA (8 criteria,
scored by automated evaluator)
==========

Your layout will be scored on 8 criteria. Higher score =
better layout. The evaluator gives ONLY numerical scores
(no text feedback), so follow these hints:

CRITICAL (highest weight): • c3_walkways [1-4]: Wide
unblocked passages between furniture. 4=all paths
comfortable. 3=main paths free, one narrow spot. 2=too
narrow. 1=blocked. • c8_entry [1-4]: Clear entry from door.
Don't block door opening. 4=wide straight passage. 3=main
path free, one spot. 2=uncomfortable. 1=fully blocked. •
c4_accessibility [1-3]: Cabinet doors/drawers open fully.
Coffee tables reachable. 3=fully usable. 2=usable with
effort. 1=blocked. Not evaluated for: bathroom, corridor,
hallway, wardrobe room.

HIGH PRIORITY: • c1_zones [1-4]: Clear functional zones
(group of 2+ items with shared function). 4=clearly
separated. 3=distinguishable but overlap. 2=hinted, blurry.
1=no zones. • c2_focus [1-4]: Seating faces focal points.
TV centred relative to sofa/chairs. 4=all seats face focus.
3=main seats OK. 2=weak focus. 1=ignored. • c5_pairs
[1-2]: Furniture groups stay together (table+chairs,
bed+nightstands, sofa+table). 2=assembled nearby. 1=broken
apart.

IMPORTANT: • c6_balance [1-3]: Fill 60-70% of room. Spread
evenly, avoid dead zones. 3=ideal. 2=partial. 1=overcrowded
or too empty. • c7_light [1-3]: Desks and seating near
windows. TV not facing windows. 3=key zones well-lit, no
glare. 2=some daylight. 1=dark or glare. Only evaluated if
room has windows.

PRIORITY ORDER: CRITICAL > HIGH > IMPORTANT

STRICT CONSTRAINTS YOUR TASK: REFINE this existing layout
to maximise the evaluation scores. You receive a starting
layout — improve it, don't redesign from scratch.

DO: widen passages, clear entry, group related furniture,
orient seating toward focal points, move furniture closer to
walls, rotate/reposition to improve zones. You may remove
at most 1 item (only if clearly justified).

DON'T: remove more than 1 original item; redesign from
scratch; place furniture outside room boundaries; overlap
non-exempt items.

WHAT CAUSES LOW SCORES: Invalid JSON or empty output →
worst possible score. Furniture outside room contour or
overlapping → score penalty. Blocked entry door or narrow
walkways → heavily penalised (CRITICAL).

==========
{'room_type': 'bathroom', 'room_vertices':
[[146,973],[1378,973],[1378,1252],[1861,1252],[1862,0],[148,3]],
'furniture': [{'object': 'bathtub', 'points':
[[0,392],[864,392],[864,0],[0,1]]}, {'object': 'toilet',
'points': [[1,197],[304,197],[303,0],[0,0]]}, {'object':
'sink', 'points': [[316,193],[316,0],[0,0],[1,193]]},
{'object': 'washing machine', 'points':
[[0,281],[0,0],[300,0],[301,278]]}], 'doors':
[{'object': 'door', 'bbox_points':
[[1862,939],[1904,939],[1904,527],[1862,527]]}], 'windows':
[]}
<!- METADATA: sample_1795 ->
```

E GRPO Per-Step Algorithm

This section expands the high-level description in §3.3 into the exact per-step procedure used in our implementation. The prompt format for each step is in §D (Stage 3 GRPO prompt).

Per-step algorithm. Step 1: sample a room prompt. Step 2: generate $K{=}4$ candidate JSON layouts; layouts that fail JSON parsing receive $r{=}0$ and skip the critic. Step 3: validate each layout by filtering malformed polygon items and checking the furniture-removal constraint (≤ 1 item may be removed). Step 4: compute structural penalties—$p_{\text{ov}}{=}0.05$ per furniture overlap, $p_{\text{bd}}{=}0.08$ per out-of-boundary item; if total violations exceed 3, skip the critic and assign $r{=}0.05$. Step 5: render valid layouts to $800{\times}600$ px images and score with the critic API on all eight criteria concurrently, obtaining raw scores $s_c^{(k)} \in \{1, \ldots, K_c\}$. Step 6: aggregate critic reward and apply sharpening:

$$r_{\text{critic}}^{(k)} = \frac{\sum_{c \in \mathcal{A}(\tau)} w_c \left(s_c^{(k)} - 1\right)}{\sum_{c \in \mathcal{A}(\tau)} w_c \left(K_c - 1\right)}, \tag{16}$$

$$r^{(k)} = \text{clip}\Big(\big(r_{\text{critic}}^{(k)} - p^{(k)} - 0.5\big) \cdot \gamma + 0.5,\ 0,\ 1\Big),$$

where $p^{(k)}$ is the total structural penalty for candidate k and $\gamma{=}1.5$ is the sharpness factor that amplifies score differences for a stronger GRPO advantage signal. Step 7: compute within-group advantages $A^{(k)}{=}r^{(k)}{-}\bar{r}$, $\bar{r}{=}K^{-1}\sum_k r^{(k)}$. Step 8: clipped-ratio GRPO update with KL penalty $\beta{=}0.01$ (Equation (3)).

F Extended Results

Stage 2 (Critic) was evaluated across 36 VLM configurations in two modes: per-criterion (18 models) and aggregated (18 models). The per-criterion full breakdown appears in the main text (Table 4, §4.2). Here we give the full aggregated-strategy table so that all configurations are documented in one place.

Table 14 ranks all 18 model configurations under the aggregated evaluation strategy (all eight criteria scored in one call). The table reveals a stark contrast with per-criterion results: LoRA

fine-tuning yields only modest gains of $+9\,\%$ QWK and $-10\,\%$ MAE, compared to $+57\,\%$ QWK and $-48\,\%$ MAE under per-criterion evaluation. Notably, the zero-shot Qwen3-VL-32B baseline (QWK = 0.420) nearly matches the best LoRA model Qwen3-VL-8B r64 (QWK = 0.458), implying that large models already internalize aggregated rubric-following from pretraining, leaving little room for LoRA to improve. An unexpected inversion appears in the size–quality relationship: the 8B LoRA model (QWK = 0.458) outperforms the 32B LoRA (QWK = 0.421), suggesting that over-parameterization impairs multi-task generalization when all criteria must be handled jointly. The anomalously low METEOR scores for Qwen2.5-VL-3B (0.108, both LoRA and baseline) indicate that 3B models produce degenerate text when asked to generate eight comments simultaneously, reverting to repetitive short phrases. These findings collectively support the recommendation in the main text: use per-criterion evaluation whenever quality matters, and consider aggregated only when inference budget is the primary constraint.

G Extended Critic Analysis

G.1 Qualitative critique example

Figure 9 shows a representative per-criterion critique on a held-out open-concept living room (Qwen3-VL-32B, LoRA $r{=}16$). The model correctly rewards the bar counter + bar stool pairing (C5 2/2) and adequate entry clearance (C8 3/4), while penalising the missing focal element in the sleeping zone (C2 3/4) and the bed placed far from both windows (C7 2/3) — each with a concrete fix suggestion.

G.2 Effect Size Analysis

Effect sizes (Cohen's d) quantify the practical significance of performance gaps beyond statistical significance alone. Table 15 reports three pairwise comparisons that together characterise the value of LoRA and the value of per-criterion evaluation. Each row shows the mean QWK difference (ΔQWK), the pooled Cohen's d, and an interpretation using the Cohen–Sawilowsky scale (<0.2 negligible, 0.2–0.5 small, 0.5–0.8 medium, 0.8–2.0 large, ≥ 2.0 very large). *Row 1* (green) compares per-criterion LoRA to per-criterion zero-shot and represents the core claim that fine-tuning is essential for this task. *Row 2* (yellow) makes the same comparison under the aggregated strategy, showing that zero-shot models are already competitive. *Row 3* (blue) compares the two best LoRA models head-to-head, directly answering whether the $8\times$ inference cost of per-criterion evaluation is justified.

The $d{=}2.41$ result means a randomly selected LoRA model beats a randomly selected baseline more than $99\,\%$ of the time. Conversely, $d{=}0.38$ for aggregated implies zero-shot baselines are sufficient for budget-constrained deployments. The $d{=}2.05$ for per-criterion versus aggregated LoRA quantifies that the $8\times$ inference overhead is strongly justified for quality-critical applications.

G.3 Error Pattern Analysis

Table 16 breaks down prediction errors for the best critic model (Qwen3-VL-32B per-criterion) into five mutually exclusive categories: exact hits, over-predictions (predicted score higher than expert), underpredictions (lower), off-by-one errors, and major errors ($|\Delta|{\geq}2$). The table has a single data row because only the best model is included; the per-column interpretation is as follows. *Exact (72.3 %)* is the fraction of criterion–image pairs where the predicted ordinal score exactly matches the expert label—a strong result for a 4-point scale. *Over-prediction (24.5 %)* and *Underprediction (3.2 %)* reveal an 8:1 asymmetric optimism bias: the model systematically tends to rate layouts slightly higher than experts, likely because the training corpus contains layouts that are good-but-not-perfect and the model learns an optimistic prior. *Off-by-one (20.3 %)* counts all $|\Delta|{=}1$ errors, regardless of direction; combined with exact, this gives $92.5\,\%$ adjacent-or-exact accuracy. *Major (7.5 %)* counts $|\Delta|{\geq}2$ errors, which represent genuinely wrong assessments and are the quality floor for production deployment.

Figure 10 provides two complementary visualisations. The *left panel* (confusion matrix) plots the $4{\times}4$ predicted vs. expert score matrix for the best model, confirming the strong diagonal and the systematic upward shift: most errors concentrate in the upper-triangle cells (prediction $>$ expert), particularly at true score 2 being predicted as 3. The *right panel* shows QWK as a function of model size (3B, 4B, 7B, 8B, 11B, 14B, 30B, 32B) for both evaluation strategies separately. For per-criterion, QWK increases monotonically with size, confirming that specialised ergonomic knowledge scales with model capacity. For aggregated evaluation, the curve is non-monotonic: the 8B model achieves the

Table 14: **Aggregated evaluation (18 models, 4 metrics).** All criteria are evaluated simultaneously in one call per layout. LoRA fine-tuning yields only modest gains (+9% QWK, −10% MAE); the zero-shot 32B baseline (QWK = 0.420) is nearly competitive with the best LoRA (QWK = 0.458).

Type	Model	B	MAE ↓	QWK ↑	Spearman ↑	METEOR ↑
LoRA fine-tuned (9 configurations)						
LoRA	**Qwen3-VL-8B** (r=64)	8	**0.5706**	**0.4579**	**0.4951**	0.2804
LoRA	Qwen3-VL-32B (r=16)	32	0.6381	0.4205	0.4577	**0.2843**
LoRA	Qwen3-VL-4B (r=16)	4	0.6237	0.4056	0.4300	0.2406
LoRA	InternVL3.5-8B (r=16)	8	0.6712	0.4044	0.4359	0.2454
LoRA	Qwen2.5-VL-32B (r=16)	32	0.6940	0.3982	0.4511	0.2781
LoRA	Qwen2.5-VL-3B (r=16)	3	0.6375	0.3920	0.4246	0.1084
LoRA	Qwen2.5-VL-7B (r=16)	7	0.6281	0.3912	0.4229	0.2667
LoRA	InternVL3.5-4B (r=16)	4	0.6988	0.3462	0.3689	0.2284
LoRA	Qwen3-VL-8B (r=16)	8	0.6955	0.3281	0.3671	0.2645
Zero-shot baseline (9 configurations)						
Baseline	**Qwen3-VL-32B**	32	0.6347	**0.4199**	**0.4585**	**0.2791**
Baseline	Qwen3-VL-4B	4	0.6303	0.4055	0.4354	0.2083
Baseline	Qwen2.5-VL-7B	7	**0.6227**	0.3983	0.4325	0.2593
Baseline	Qwen2.5-VL-3B	3	0.6350	0.3947	0.4267	0.1080
Baseline	Qwen2.5-VL-32B	32	0.7111	0.3921	0.4473	0.2767
Baseline	InternVL3.5-8B	8	0.6862	0.3722	0.4127	0.2652
Baseline	Qwen3-VL-8B (run 1)	8	0.6979	0.3477	0.3830	0.2597
Baseline	Qwen3-VL-8B (run 2)	8	0.6899	0.3461	0.3861	0.2603
Baseline	InternVL3.5-4B	4	0.7552	0.3138	0.3427	0.1909

Table 15: **Effect size analysis (Cohen's d).** Per-criterion LoRA fine-tuning is essential (d=2.41, very large); aggregated LoRA provides only marginal gains (d=0.38, small–medium). Scale: <0.2 negligible, 0.2–0.5 small, 0.5–0.8 medium, 0.8–2.0 large, ≥2.0 very large. PC = Per-Criterion, AG = Aggregated.

Comparison	ΔQWK	Cohen's d	Effect magnitude
PC-LoRA vs. PC-Baseline	+0.242	**2.41**	Very large (LoRA essential)
AG-LoRA vs. AG-Baseline	+0.038	0.38	Small–medium (LoRA optional)
PC-LoRA vs. AG-LoRA	+0.205	**2.05**	Very large (8× cost justified)

Table 16: **Error pattern analysis — best critic (Qwen3-VL-32B, LoRA r=16, per-criterion).** Predicted scores compared to expert scores on the held-out set; Δ is the signed difference. An 8:1 overprediction-to-underprediction ratio reveals a systematic optimism bias amenable to post-hoc calibration; only 7.5 % of predictions are catastrophic (|Δ|≥2).

Error type	Definition	Share (%)
Exact	$\Delta = 0$	**72.3**
Overprediction	$\Delta > 0$	24.5
Underprediction	$\Delta < 0$	3.2
Off-by-one	$\lvert\Delta\rvert = 1$	20.3
Major	$\lvert\Delta\rvert \geq 2$	7.5

best QWK (0.458), and the 32B model underperforms (0.421), providing visual evidence for the over-parameterization hypothesis discussed above.

The 72.3 % exact accuracy (from Table 16) confirms that QWK = 0.663 is not inflated by lenient ordinal scoring. Combining exact and off-by-one gives 92.5 % adjacent-or-exact, matching the human inter-annotator ceiling of 92.3 %. The 8:1 overprediction bias (left panel of Figure 10) is an optimism bias, correctable via temperature or Platt scaling without retraining; only 7.5 % of predictions are catastrophic (|Δ|≥2), which is acceptable for production use with confidence-based human review.

G.4 Per-Criterion Model Family and Size Scaling

For per-criterion evaluation, model family hierarchy is Qwen3-VL > Qwen2.5-VL > InternVL3.5 across all sizes; size scaling follows 32B > 8B >

4B > 3B (right panel of Figure 10, Figure 10b), indicating that specialised ergonomic knowledge benefits from model capacity. LoRA rank r=16 versus r=64 gives QWK 0.646 versus 0.649 (+0.5 %) at 4× more parameters, making r=16 production-optimal.

The aggregated strategy shows a notable inversion: the 8B model (r64) outperforms the 32B model on QWK (0.458 vs. 0.421), suggesting over-parameterization impairs multi-task generalization. Qwen2.5-VL-3B shows anomalously low METEOR (0.108) in both LoRA and baseline aggregated conditions, indicating the 3B model produces degenerate text outputs under combined multi-criterion prompting (full results in Table 14).

H Qualitative Examples and Extended Related Work

H.1 Extended Related Work

Pre-LLM indoor scene synthesis. ATISS (Paschalidou et al., 2021) introduced autoregressive Transformers for furniture arrangement on rectangular floor plans, establishing the paradigm LLM methods build upon. HouseDiffusion (Shabani et al., 2023) generates polygon-format floorplans via diffusion, but both works use axis-aligned rooms with no ergonomic quality signal.

LLM/VLM-based layout generation. LayoutGPT (Feng et al., 2023) prompts LLMs with stylesheet representations for 2D/3D spatial layouts using in-context exemplars. LLplace (Yang et al., 2024) fine-tunes Llama3 on 3D-Front dialogue for interactive bounding-box layout in convex rooms. OptiScene (Yang et al., 2025) applies two-stage SFT then DPO (Rafailov et al., 2023) to align layouts with human preferences, but uses offline pairwise labels rather than an automated per-criterion GRPO reward. I-Design (Çelen et al., 2025) uses a multi-agent LLM and scene graph without structured ergonomic scores. Chat2Layout (Wang et al., 2024) refines convex 3D layouts via visual feedback, without a learned multi-criterion critic. DirectLayout (Ran et al., 2025) uses BEV planning with a dual VLM/LLM reward, conceptually close to Stage 3 but confined to rectangular rooms and bounding boxes. StructLayoutFormer (Hu et al., 2025) separates structural from positional predictions, motivating our polygon output; ReSpace (Bucher and Armeni, 2025) applies SFT and preference alignment but stays in rectangular rooms.

Spatial reasoning and evaluation. Spatial 3D-LLM (Wang et al., 2025b) shows VLMs can learn structured spatial reasoning, a capability we exploit via per-criterion LoRA fine-tuning. SceneEval (Tam et al., 2025) checks object presence and collisions but not ergonomic quality. LayoutVLM (Sun et al., 2025) uses a VLM for physical plausibility without a human-aligned ergonomic critic.

Positioning FURLA. All reproducible prior methods assume rectangular rooms or convex layouts (Paschalidou et al., 2021; Yang et al., 2024, 2025; Bucher and Armeni, 2025; Ran et al., 2025; Wang et al., 2024), and measure quality via FID, collision rates, or pairwise labels—none capture multi-dimensional ergonomic quality with inter-annotator validation. FURLA introduces an eight-criterion rubric validated by two independent experts (Cohen's κ=0.515) and trains a LoRA VLM critic reaching QWK = 0.663 (d=2.41 over zero-shot, very large effect). Unlike OptiScene, ReSpace, or DirectLayout, which rely on offline or supervised reward pipelines, FURLA refines furniture layout via an online GRPO loop, harvesting critic rewards for K=4 candidates per step with no human annotation at inference. All prior methods output bounding boxes; FURLA generates full polygon vertex coordinates, enabling exact wall-contact placement and per-criterion door and window clearance reasoning.

I Licenses and Intended Use

Used artifacts. We use the Qwen3-VL (Bai et al., 2025a) and Qwen2.5-VL (Bai et al., 2025b) model families under the Apache-2.0 License, and InternVL3.5 (Wang et al., 2025a) under the Apache-2.0 License. Our use of all backbones is limited to model fine-tuning and inference for research purposes, which is consistent with the intended use stated by each publisher. Training frameworks are released under permissive open-source licenses: LLaMA-Factory (Zheng et al., 2024), DeepSpeed (Rasley et al., 2020), and TRL (von Werra et al., 2020) (Apache-2.0); FlashAttention-2 (Dao, 2023) (BSD-3-Clause). All training and evaluation usage follows the research-purpose scope each licence permits.

Released artifacts. The 53,332-room dataset, the 1,121 expert-annotated layout images with eight-criterion scores, the trained LoRA adapters (Stage 1 generator and Stage 2 critic), and the GRPO checkpoint will be released for non-commercial research use under a CC BY-NC-SA 4.0 license. Floor plans were collected from publicly accessible real-estate developer pages; we redistribute only the anonymised polygon vertex representations and labels, not the original raster images, and the release respects fair-use limits for research datasets. During conversion, addresses, developer metadata, owner or tenant names, natural-language listing text, and raster text are discarded; the released records contain only room polygons, door/window boxes, furniture categories, placed furniture polygons, and rubric scores. Any derivative use of these artifacts must remain within research contexts, consistent with the access conditions of the source pages.

J Future Work and Planned Improvements

Dedicated critic for geometric violations. A separate VLM critic trained to detect and score furniture intersections and out-of-boundary placements would complement the ergonomic critic, providing a learned signal for geometric validity instead of hand-crafted structural penalties. We plan to collect annotations for such violations and integrate a geometry-focused critic into the GRPO reward.

3D export to Revit. Extending the pipeline to 3D by lifting polygon layouts into Autodesk Revit (`.rvt` / Revit API or IFC) would enable full BIM workflows: furniture families with heights, material finishes, and daylight simulations directly in the designer's toolchain.

Advanced RL training algorithms. GRPO is one representative of a rapidly evolving family of group-based policy optimisation methods. Recent alternatives—DAPO (Decoupled Clip and Dynamic Sampling Policy Optimisation), RLOO (REINFORCE Leave-One-Out), and Dr. GRPO—address known instabilities such as entropy collapse and advantage distribution shift. Integrating these algorithms into the FURLA training loop is a natural next step, as they have demonstrated more stable training dynamics and higher final reward on reasoning tasks with complex multi-dimensional reward signals, directly analogous to our eight-criterion ergonomic score.

Building-code and regulatory compliance. A natural extension of the critic rubric is to add a compliance criterion that checks generated layouts against normative documents: fire-egress clearances, accessibility standards (*e.g.*, ADA, ISO 21542), minimum aisle widths, and sanitary norms. Such a criterion would turn FURLA into a design-assist tool that not only optimises ergonomic comfort but also flags regulatory violations at generation time, without requiring a separate compliance-checking pipeline.

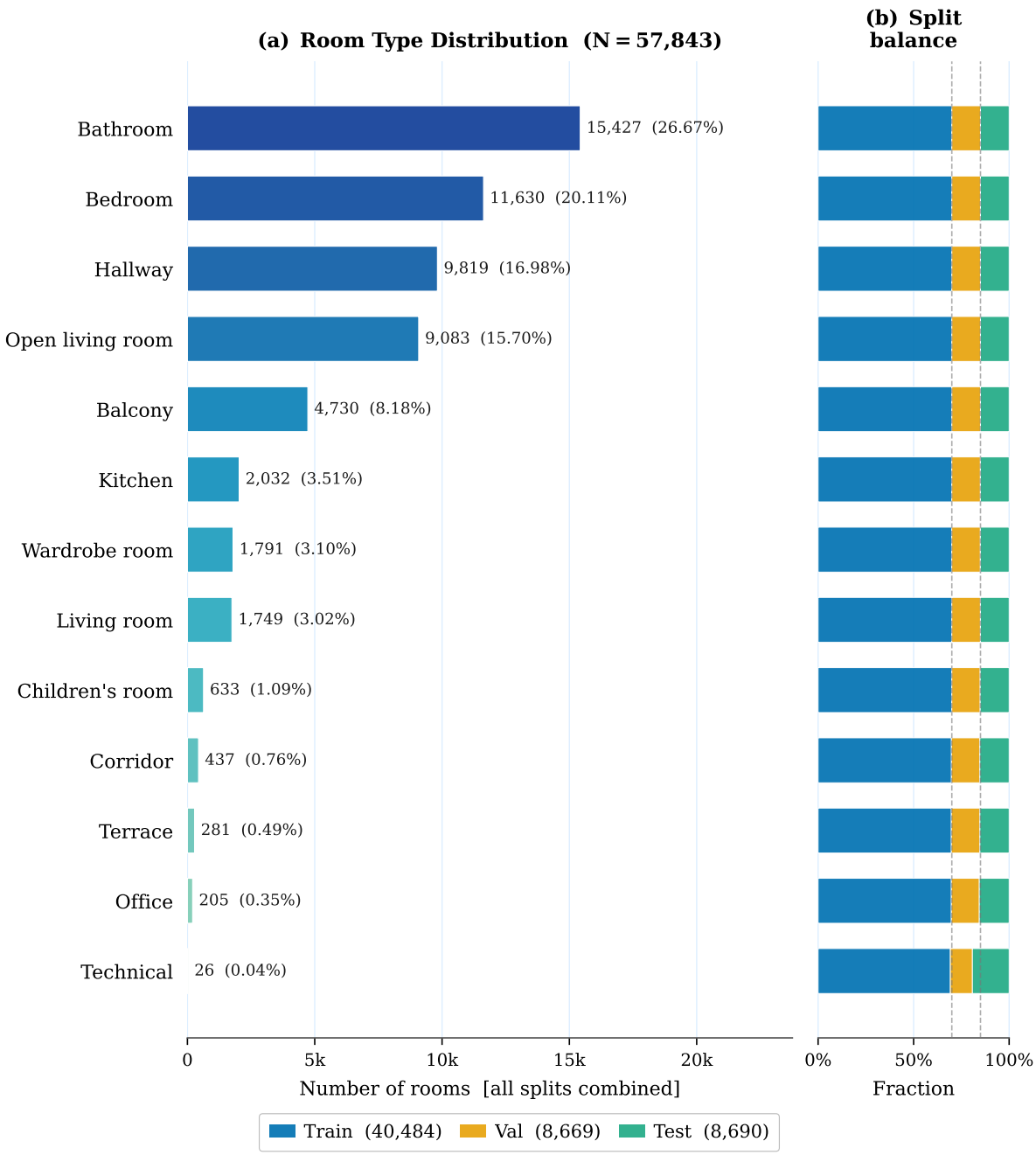


Figure 6: **Stage 1 dataset overview.** 53,332 rooms with irregular polygonal boundaries from 22,000 real apartments, spanning 13 room types and 45 furniture categories.

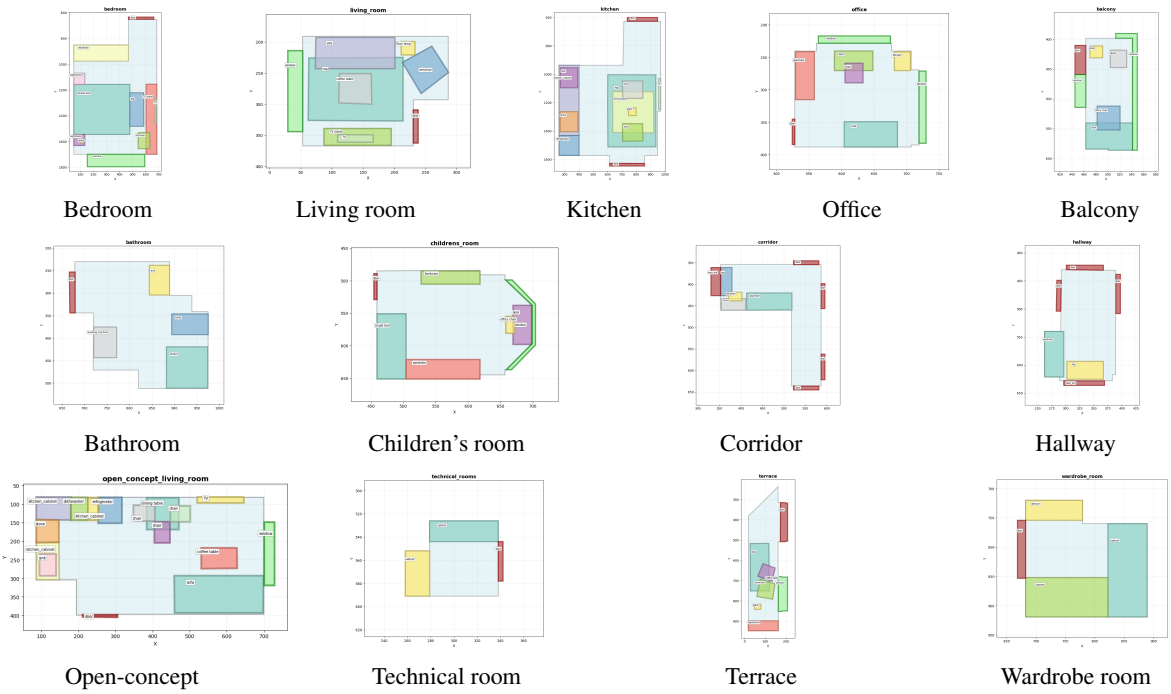


Figure 7: **Stage 1 dataset: annotated room examples across all 13 room types.** Each image shows a rendered floor plan with placed furniture polygons. All 13 room types in the 53,332-room dataset are represented.

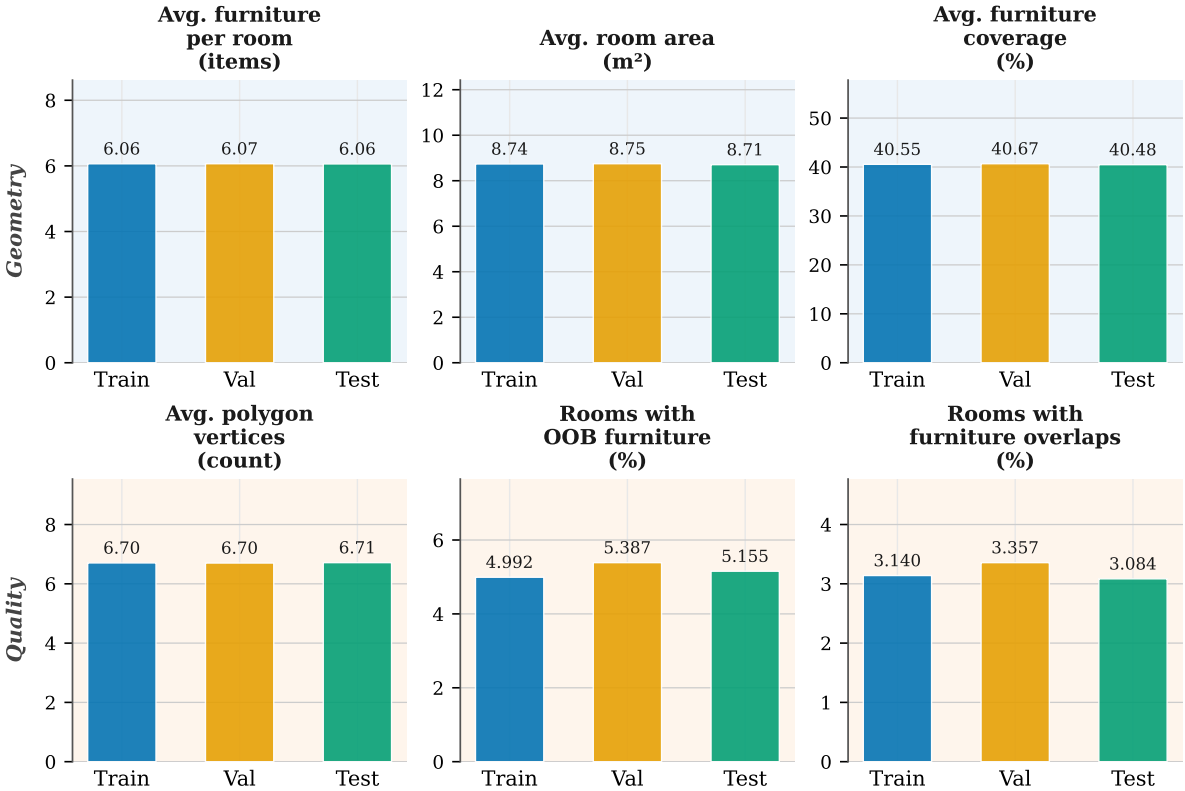


Figure 8: **Stage 1 dataset split statistics.** Train/val/test splits are well-balanced across all geometry metrics. Top row: average furniture items per room ($\approx$6.1), average room area ($\approx$8.7 m^2), and furniture coverage ($\approx$40.6 %). Bottom row: average polygon vertex count ($\approx$6.7), rooms with out-of-boundary furniture ($\approx$5 %), and rooms with furniture overlaps ($\approx$3 %). The near-identical values across splits confirm unbiased partitioning.

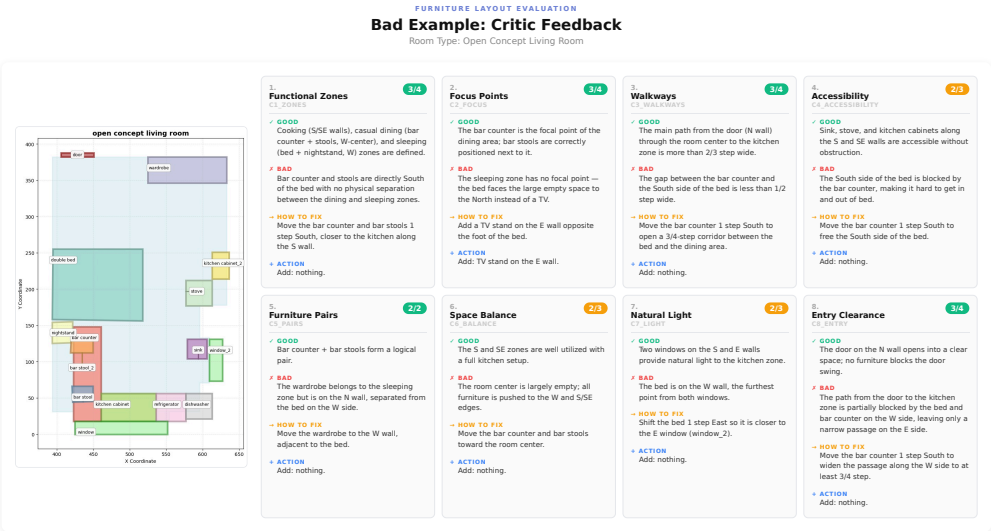
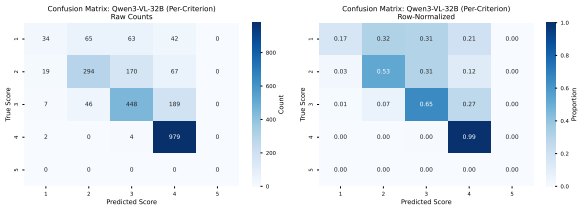
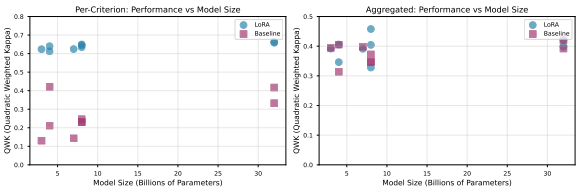


Figure 9: **Stage 2 per-criterion critique (Qwen3-VL-32B, LoRA r=16).** Open-concept living room 8337: floor plan (left) and eight-criterion scores with fix suggestions (right). The model rewards furniture pairing and entry clearance, and flags the missing focal element and poor natural light for the bed.



(a) Confusion matrix: strong diagonal (72.3%) with 8:1 over-prediction bias.



(b) QWK vs. model size: monotonic scaling for per-criterion, non-monotonic (8B optimal) for aggregated.

Figure 10: **Error structure and size scaling.** Errors are structured rather than random and are amenable to calibration. Model-size optimality depends on the evaluation strategy.

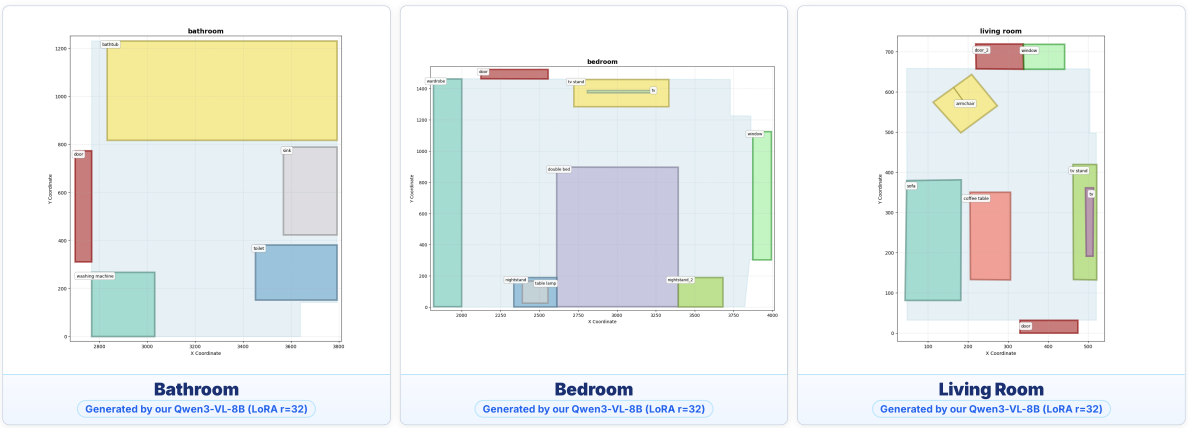


Figure 11: **Stage 3 GRPO generation examples (Qwen3-VL-8B, LoRA r=32).** Three held-out test rooms—bathroom, bedroom, and living room—generated by our best-performing model after Stage 3 GRPO reward optimisation.