


Through the PRISM: Principle-Aware, Interpretable, and Multi-Scale Evaluation of Visual Designs

Mona Gandhi^{1*}, K J Joseph², Srinivasan Parthasarathy¹, Sayan Nag²

¹Ohio State University, ²Adobe Research

{gandhi.255, srini}@osu.edu, {josephkj, snag}@adobe.com

Abstract

Effective visual communication stems from the harmony of multiple design principles, such as readability, contrast, alignment, overlap, and coherence, which collectively govern clarity and intent of the communicator. While human designers reason holistically over these principles, machine agents typically condense them into a single heuristic score, offering limited interpretability and diagnostic precision. To address this gap, we introduce  **PRISM** (**PR**inciple-aware, **I**nterpretable, and **S**tructure-guided Design Modifications), a benchmark that systematically perturbs professional layouts from the Crello dataset along measurable design principles. The benchmark comprises **100K** perturbed training samples and **10K** perturbed validation designs, each isolating a specific principle violation for controlled analysis of multimodal reasoning about design quality. We show that models like *Qwen-2.5-VL* and *GPT-4o-mini* are largely insensitive to targeted principle degradations, whereas *GPT-4o* exhibits global awareness without fine-grained disentanglement. Building on these insights, we propose a multi-scale evaluation framework that integrates lightweight scorers for quantitative assessment, instruction-tuned vision-language models for localised feedback, and prompt-based methods for global reasoning. Our framework provides interpretable explanations of design failures. Using these localised insights, we show targeted refinements that improve layout quality. Together, **PRISM** and our framework lay the foundation for interpretable design-literate multimodal reasoning systems.

1. Introduction

Designing is inherently a process of reasoning about perception and communication. When evaluating a visual composition, human designers do not typically rely on a single notion of “good design”. They often weigh multiple interacting principles such as *coherence*, *readability*, *contrast*,

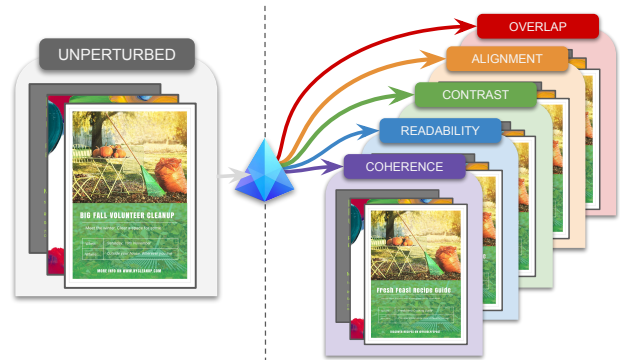



Figure 1. We introduce  **PRISM**, a benchmark that operationalizes design evaluation through measurable principles of visual composition. Derived from professional layouts, PRISM systematically perturbs examples along five principles: *coherence*, *readability*, *contrast*, *alignment*, and *overlap*, isolating targeted violations while preserving other attributes. These controlled perturbations enable interpretable and principle-specific evaluation of models’ sensitivity to localised design degradations and their holistic reasoning about visual quality.

alignment, *balance*, and many more such design principles [46]. A designer might praise a layout’s overall theme yet still adjust text spacing for legibility or modify color contrast for accessibility. This layered reasoning, which moves fluidly between global intent and local detail, enables designers to make targeted refinements preserving the composition’s purpose while enhancing clarity and aesthetics [19].

In contrast, existing computational evaluation methods reduce this multifaceted process to a single scalar score, such as aesthetic preference or layout similarity [13, 35, 38]. While convenient, these metrics cannot reveal which aspects of a design are successful or unsuccessful, and they offer little guidance for actionable improvement [30, 59] or provide end-to-end refiners [13]. Their aggregation of all principles into a single scalar is at odds with how human designers critique and refine, focusing on specific violations that can be iteratively addressed [30, 56].

Recent large language and multimodal models (LLMs

*Work done at Adobe Research

and MLLMs) can produce free-form critiques that appear similar to human reasoning [10, 30, 56]. However, their evaluations are stochastic, qualitative, and non-reproducible. Given the same poster, such models often generate inconsistent feedback, sometimes suggesting unnecessary global edits for well-designed layouts. Further, their responses lack grounding in explicit design principles and provide judgments without measurable justification. Thus, they behave more like subjective critics than consistent evaluators.

To bridge this gap, we introduce **PRISM** (PRinciple-aware, Interpretable, and Structure-guided Design Modifications), a curated perturbation dataset that systematically degrades professionally crafted designs from the Crello corpus [52] along controlled dimensions of *coherence*, *readability*, *contrast*, *alignment*, and *overlap* (ref Fig. 1). Each perturbation isolates a targeted violation, for example, text color blending into the background for readability, or a mismatch between the headline and imagery for coherence, thus providing principle-specific supervision for multimodal evaluation. These controlled degradations reflect how humans perceive drops in design quality: localised changes in text or contrast disrupt readability, while global inconsistencies undermine coherence and intent. Analysis on PRISM reveals a gap in current models: architectures like Qwen-2.5-VL [1] and GPT-4o-mini [33] often fail to register these controlled degradations, while GPT-4o [20] shows broad semantic sensitivity but lacks consistent separation between fine-grained principles.

Building on PRISM, we propose a structured evaluation framework with principle-specific **scorers** and **localisers**. PRISM-scorer is a lightweight classifier built on a SigLIP-v2 [49] backbone fine-tuned on design data for each principle, trained to produce quantitative scores. For local principles such as readability, we pair these scorers with an instruction-tuned Qwen-2.5-VL model that not only detects issues but also explains where they occur, similar to how a designer highlights specific regions needing adjustment. For higher-level properties such as coherence, which involve semantic and thematic consistency, we use prompting-based evaluation with MLLMs. Finally, we illustrate how the scorer and localiser enable interpretable, iterative refinement using a lightweight **beam search-based editing** [44]. The editor proposes minimal adjustments that improve targeted principle scores while preserving the overall layout, offering controllable feedback beyond generative MLLMs [7, 20] that treat designs as static images.

Our work makes three key contributions. We

- Present **PRISM**, a controlled perturbation dataset that isolates principle-specific degradations in designs.
- Propose a principled evaluation framework where each principle has a scorer for quantitative assessment and a localiser for spatially grounded feedback.

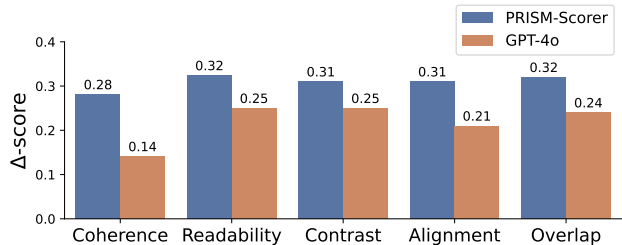


Figure 2. **PRISM-Scorer Performance.** Δ -score bar plot comparing the sensitivity of different models to PRISM perturbations. Each axis shows the change in principle-specific scores between the original and perturbed posters, as assigned by each scorer. Higher values indicate stronger discrimination of principle violations. Refer to Section 4.3 for further details.

- Demonstrate that our SigLIP-based scorers cleanly separate principle-specific degradations, outperforming GPT-4o in disentanglement (See Fig.2), and that our instruction-tuned Qwen-2.5-VL localiser aligns most closely with ground-truth annotations for local principles. These advances establish a principled and interpretable pipeline for evaluating and refining visual design.

2. Related Works

Design Evaluation. Design evaluation has been extensively studied across domains and development stages. Early heuristic-based methods [2, 15, 31, 37, 54] struggled to capture semantic design attributes. Subsequent approaches [9] enhanced evaluation capabilities but lacked comparative scoring mechanisms, motivating siamese-based approaches [13, 57]. Kong et al. [23] trained a model for aesthetic assessment, while Tabata et al. [48] perturbed layouts to synthesize poor designs for scoring. More recently, MLLM-based methods have focused on design generation [22], assessed only limited attributes [14, 22], or provided actionable feedback [10, 30]. However, such feedback is often stochastic, redundant, over-suggestive, or lacks contextual awareness - preventing principled, quantifiable and localised assessment of crucial design attributes.

Design Principles. Core principles of graphic layout were established long ago and remain widely used [32, 37]. Since then they have been used for guiding layout generation [28, 45]. Recently, layout-generation methods [16, 17, 25, 27, 56] have adopted content based metrics: occlusion, unreadability and graphic metrics including overlay (IoU), non-alignment, using them to iteratively refine layouts toward structurally and visually plausible compositions. Other works also leverage design-related cues for downstream tasks, for example DocEdit-v2 [47] incorporates spatial coherence in multimodal document editing, and hybrid global-local segmentation models use spatial

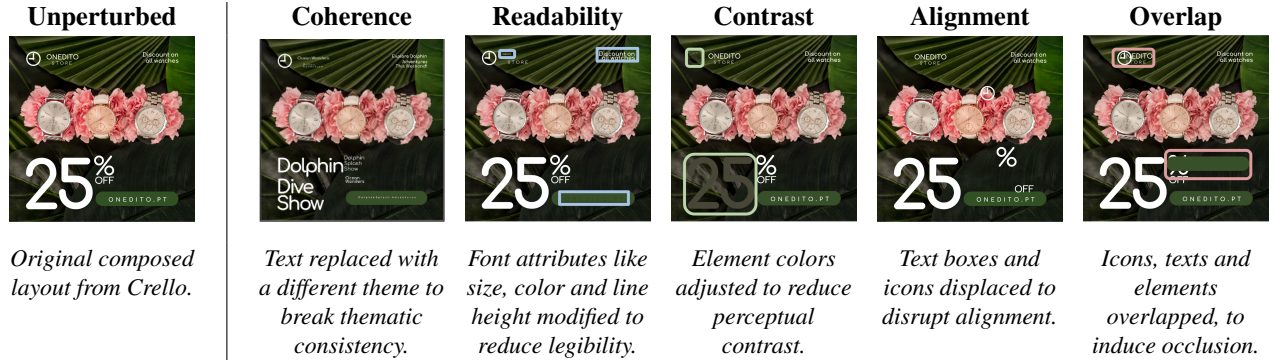


Figure 3. **Illustration of PRISM perturbations.** The composed layout (left) serves as the reference design. Each perturbed layout (right) isolates a violation of a specific design principle: coherence, alignment, readability, contrast, and overlap, while minimally affecting others. Descriptions below each image summarize the targeted modification. Bounding boxes highlight regions where local edits were applied, providing ground-truth supervision for principle-specific localisation.

coherence and region–text alignment for grounding [51]. Even outside visual design, multimodal topic-modeling introduces metrics to capture visual coherence and topic separability [12]. Across these directions, design principles consistently operate as guidance signals for generation and multimodal reasoning, but they remain largely unused as explicit dimensions for assessing design quality.

Visual grounding with MLLMs. Recent MLLMs [3, 6, 24, 39, 40, 42, 50, 53, 58] have substantially advanced the integration of visual and linguistic understanding, achieving impressive performance in tasks such as referring expression comprehension and segmentation. These models showcase the growing capability of LLMs to perform fine-grained localisation and complex reasoning over visual content. Approaches such as [3, 39, 50] emphasize language-driven grounding by constructing rich contextual representations. Whereas methods like [24, 40, 42, 42, 58] leverage joint vision-language embeddings to generate segmentation masks. Further, [5, 29] explore multi-modal grounding by incorporating audio-visual cues for spatial and temporal understanding. Despite these advances, visual grounding within MLLMs remains largely confined to image- and video-based applications and has yet to be extended to specialized domains such as defect localisation in designs.

3. Principle-Aware Design Evaluation

To enable interpretable and principle-specific evaluation of visual design quality, we introduce a **principle-aware evaluation framework** built upon a curated dataset of controlled layout perturbations. Section 3.1 presents **PRISM**, a principle-disentangled dataset derived from professional designs that isolates key visual composition principles through targeted modifications. Section 3.2 then describes the **scorer** and **localiser** modules that leverage PRISM to quantitatively and spatially assess design violations.

3.1. **PRISM: Principle-Disentangled Dataset**

To enable interpretable and principle-specific evaluation of visual layouts, we introduce **PRISM (PRinciple-aware, Interpretable, and Structure-guided Design Modifications)**, a curated dataset that disentangles five fundamental principles of design quality: *coherence*, *readability*, *contrast*, *alignment*, and *overlap* (See Fig. 1). PRISM is derived from the Crello corpus [52] of professional designs, which provides structured metadata describing each design’s geometry, text attributes, color palette, and so on. This metadata enables controlled perturbations of individual design components while preserving other aspects of composition.

Principle definitions. We define the principles as follows: *Coherence* measures semantic consistency between textual and visual elements; *Readability* measures typographic clarity and legibility; *Contrast* measures perceptual distinctness between overlapping or adjacent regions; *Alignment* measures geometric regularity of text and element placement; and *Overlap* measures unintended occlusions between visual components. These definitions serve as the organizing factors guiding the perturbations [19, 26, 46].

Perturbation methodology. For each principle $p \in \mathcal{P}$, we generate a perturbed layout $x'_p = x + \Delta_p$, where Δ_p introduces a targeted violation of p while keeping other principles approximately invariant. **Global** principles: *coherence* and *alignment* require semantic or structural modifications, whereas **local** principles: *readability*, *contrast*, and *overlap* involve perceptual or spatial adjustments. Across all principles, GPT-4o-based [20] prompting is used to modify textual content, replace icons or background imagery, and suggest RGB color values for altered text and elements. The prompt does not include the poster as a whole but only the required details for edits (Supp.A). In parallel, geometric and typographic attributes such as text-box coordinates, font size, etc are adjusted directly through the Crello meta-

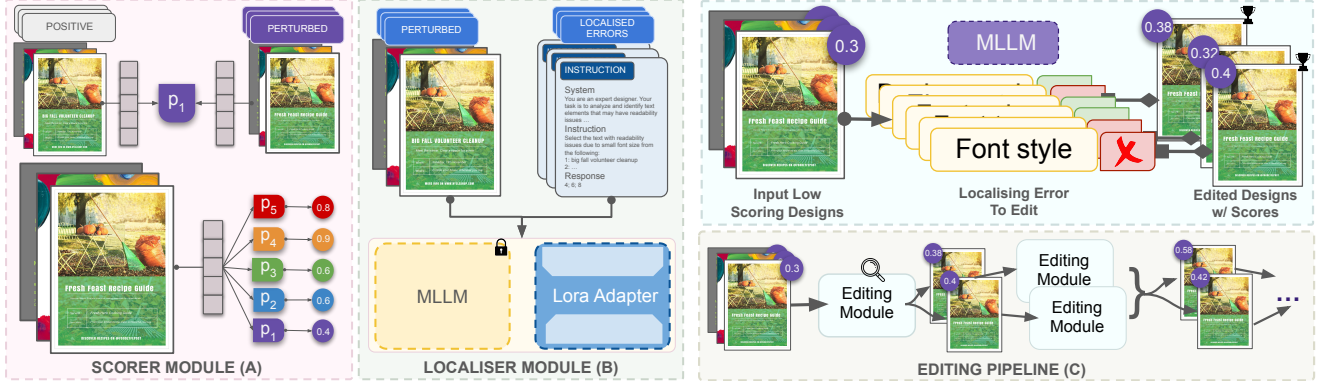


Figure 4. Using \blacktriangle PRISM, we build two modules for principle-aware design evaluation: (A) a **scorer** that quantifies quality across five design principles using controlled perturbations, and (B) a **localiser** that identifies regions responsible for readability or contrast degradations. A **demonstrative editing pipeline** (C) operates on the coherence module, where GPT-4o serves as a prompting-based localiser to identify global inconsistencies and propose edits. The pipeline employs a **beam search** strategy, generating candidate edits from the localiser and selecting the top two (beam size = 2) using the scorer. Together, these components establish a foundation for interpretable, end-to-end design assessment and refinement.

data, scaled relative to their original values to ensure realistic, layout-consistent perturbations. For local principles, we additionally record the precise set of modified regions and corresponding changes, providing structured supervision that can be used to evaluate or guide localisation-based editing techniques (See Section 3.2.2).

Example. Figure 3 illustrates representative perturbations. In the *coherence* example, the text promotes a “dolphin show” while the icons and background depict wristwatches, violating thematic consistency. In the *readability* example, we intentionally reduced the font size for the word “oned-ito”, which reduces readability.

Although complete independence among principles is inherently challenging; each perturbation is constructed to emphasize a single factor with minimal collateral impact on others. The resulting dataset comprises paired composed and perturbed layouts $\{(x, x'_p)\}$, enabling quantitative and spatially grounded evaluation of design quality. **PRISM** is derived from the Crello corpus, which provides approximately 20K training and 2K validation professional designs. Each principle is perturbed independently over this split, yielding about **20K training** and **2K validation** per principle, and totaling roughly **100K training** and **10K validation** layout pairs across all five principles.

3.2. Scorer and Localiser Modules

Building on PRISM, we develop two complementary modules for design evaluation. The **binary scorers** quantitatively assess whether each principle is satisfied in a given layout, while the **localiser** provides spatially grounded feedback identifying regions responsible for the violation. Together, these components enable fine-grained, interpretable evaluation of visual layouts and serve as inputs

for the editing pipeline described in Section 6.

3.2.1. Principle-Aware Scorer

We define a collection of *principle-aware grounding functions* $\mathcal{G} = \{g_p\}_{p \in \mathcal{P}}$ that estimate the likelihood that a visual layout satisfies specific design principles, such as *readability*, *contrast*, *alignment*, *overlap* and *coherence*. Each grounding function g_p operates over a design representation $\phi(x)$ and outputs a probability $P(p|x) = g_p(\phi(x))$, quantifying adherence to principle p . Together, these functions provide an interpretable decomposition of overall design quality, mapping a layout’s multimodal embedding space into principle-specific probabilities. This formulation enables a structured and transparent evaluation framework that moves beyond a single, opaque score toward measurable reasoning aligned with human design intuition.

Design-Aware Fine-tuning. We begin by adapting a pretrained multi-modal encoder to become *design-aware*, ensuring its visual representations capture structural and compositional cues relevant to layout evaluation. At this stage, only the vision encoder is updated, while the text encoder remains frozen. The model is trained on multi-modal document-caption pairs, where the captions convey the overall design intent of the document. Through this adaptation, the encoder learns layout embeddings that capture spatial arrangement, visual balance, and typographic hierarchy. The resulting design-aware vision encoder produces fixed embeddings $\phi(x)$ for all scorers, ensuring consistent and semantically grounded feature representations across principles.

PRISM-Scorer. We train independent binary classifiers on top of the frozen design-aware embeddings to estimate

whether each design principle $p \in \mathcal{P}$ is satisfied. Given a layout x , the encoder produces $\phi(x)$, and the corresponding scorer predicts:

$$g_p(x) = \sigma(\phi(x)^\top W_p), \quad (1)$$

where $W_p \in \mathbb{R}^d$ are the learnable weights for principle p , and $\sigma(\cdot)$ denotes the sigmoid activation. Training uses paired composed and perturbed examples, where each perturbation selectively violates a single design principle. Each scorer is optimized using binary supervision and learns to assign higher probabilities to principle-aligned layout designs and lower probabilities to those designs that are at odds with key design principles. The resulting outputs $g_p(x) \in [0, 1]$ form probabilistic, interpretable scores indicating the degree of adherence to each principle, as illustrated in Fig. 4(A). Collectively, the set of principle-specific functions $\mathcal{G}(x) = \{g_p(x)\}_{p \in \mathcal{P}}$ constitutes a structured, interpretable evaluation space that decomposes overall design quality into measurable dimensions aligned with human reasoning.

3.2.2. Principle-specific Localisation of Errors

While the binary scorers provide quantitative assessments of whether a design adheres to each principle, they do not reveal *where* violations occur. To enable spatially and semantically grounded interpretability, we introduce a principle-specific localisation module that evaluates beyond scalar scores. This module identifies the precise visual or textual elements responsible for violating a given design principle, bridging quantitative scoring with qualitative explanation. Localisation is modeled at two complementary levels: (i) a **localised** component that detects fine-grained perceptual issues such as readability, contrast, and overlap, and (ii) a **global** component that reasons about higher-level semantic properties such as coherence and alignment.

Localised Reasoning for Perceptual Principles. To localise design violations and provide interpretable feedback for perceptual principles such as *readability*, *contrast*, and *overlap*, we fine-tune a vision-language model using structured instruction prompts derived from localised perturbations in PRISM. This module identifies the specific *elements* within a layout that fail to satisfy a given principle.

During training, each example is formulated as an instruction-based multiple-choice task with explicit candidate options from the metadata. For *readability*, the input includes the layout image and a list of text elements, each associated with a unique element identifier and its textual content. For *contrast* and *overlap*, the input contains pairs of elements that occupy intersecting regions in the design space, also represented by their corresponding IDs. The instruction then prompts the model to select which elements or element pairs violate the specified principle, for example: “Select the text with readability issues due to small

font size.” To establish consistent reasoning behavior, we include a system prompt that defines the model’s role and context at the start of every interaction, shown in Fig. 4(B).

Given an image–instruction pair (x, c_p) and a finite candidate set $\mathcal{O} = \{o_1, o_2, \dots, o_n\}$, where each o_i denotes an element ID and its associated content, the model predicts a subset $\hat{\mathcal{O}}_p \subseteq \mathcal{O}$ corresponding to the violating elements:

$$\hat{\mathcal{O}}_p = \text{Model}(x, c_p, \mathcal{O}). \quad (2)$$

Training minimizes the cross-entropy loss between the predicted and ground-truth selections of violating element IDs.

During inference, the fine-tuned model outputs the list of IDs of the elements that violate the queried principle, without generating free-form text. This design choice constrains the output space to a discrete set of candidates, improving reliability and reproducibility over generative responses.

Global Reasoning for Semantic Principles. While perceptual principles such as *readability*, *contrast*, and *overlap* can be localised to specific regions, global principles like *coherence* and *alignment* require reasoning over the entire layout. PRISM provides structured perturbations for these principles, but a definitive ground truth is not available since multiple valid revisions can achieve coherence or alignment. We therefore use a prompt-based evaluation that checks whether major visual components align with the layout’s inferred theme and structure, enabling holistic assessment of semantic and organizational fidelity.

For instance, while examining *coherence*, the model first infers the design’s overarching theme (e.g., “Kid’s Birthday Party”) from its textual content and derives representative visual expectations for each component, such as a playful background, festive icons, and bold, cheerful typography. Each component (background, icons, font style, etc.) is then queried using structured prompts to verify alignment with these expectations. This procedure identifies components that visually differ from the inferred theme.

An analogous formulation is applied to *alignment*, where the model verifies whether elements such as text boxes or icons deviate from the underlying structural grid. The resulting qualitative judgments provide interpretable indicators of global principle violations; not all require correction, and resolving one often restores multiple inconsistencies.

4. Experimental Setup

Our experiments evaluate whether PRISM produces clean, principle-specific degradations and examine how effectively our scorer–localiser framework and contemporary MLLMs detect and respond to these controlled violations.

4.1. Validating Principle Disentanglement

To assess whether PRISM perturbations successfully isolate individual design principles, we conduct a controlled human study with 15 participants. Each participant is shown

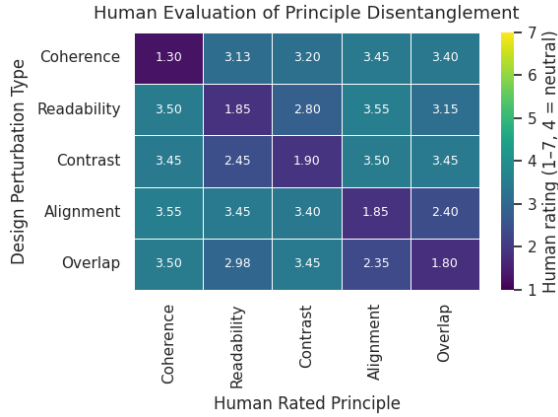


Figure 5. **Human evaluation of principle disentanglement.** Average ratings (1–7, 4 = neutral) across 10 poster pairs per principle show strong diagonal responses, confirming that PRISM perturbations primarily affect their intended design principles.

10 paired examples, where each pair contains a professionally designed poster and its perturbed variant, without indicating which is which. For every pair, three independent participants rate which design better satisfies each of the five principles (*coherence*, *readability*, *contrast*, *alignment*, and *overlap*) on a 1–7 rating scale, where 1 indicates the original is much better, 7 indicates the perturbed is much better, and 4 indicates no clear preference. We get the average of three participants for each pair and calculate the mean across all pairs. This measures how selectively each perturbation affects its intended principle.

4.2. Training of Principle-Aware Scorers

Each principle-specific scorer is trained on a design-aware SigLIP-v2 vision encoder [49], following the setup in Section 3.2.1. The encoder is adapted using the training splits of the Creatidesign [55] and Crello [52] datasets, containing 400K and 20K layout–caption pairs. Captions describing each poster’s theme and composition are generated with GPT-4o [20]. Only the vision encoder is updated in this stage, enabling it to capture structural, spatial, and typographic cues relevant to layout understanding.

Using the frozen encoder, we train five binary classifiers, one for each design principle, to distinguish between composed and perturbed layouts. Each scorer is trained and evaluated on balanced sets comprising composed layouts from the Crello dataset and their corresponding perturbations from the PRISM dataset, using 40K images for training and 4K for testing. A 90–10 split is used for validation within the training set. Training employs binary cross-entropy loss, with model selection based on validation AUC, and early stopping is applied after seven consecutive non-improving epochs. We report Precision, Recall, F1, and AUC to quantify scorer performance across principles.

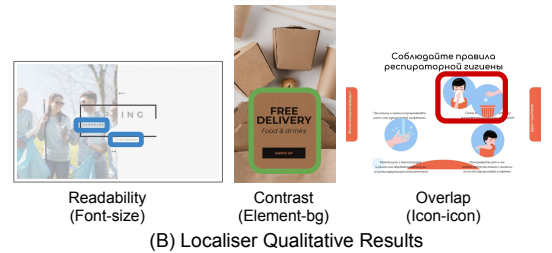


Figure 6. **Qualitative results.** (A) **PRISM-Scorer outputs.** The left image in each pair is the original layout and the right shows its PRISM perturbation, with scores indicating the predicted principle-specific degradation. (B) **Localiser outputs** from the Qwen-2.5-VL instruct-tuned model. The highlighted text specifies the targeted principle and the sub-aspect being queried. For example, in readability we prompt the model to attend specifically to small font-size regions that impair clarity.

4.3. Model Sensitivity to PRISM Perturbations

To assess how well different models capture principle-specific degradations, we evaluate four systems on PRISM: GPT-4o, GPT-4o-mini, Qwen-2.5-VL (7B) [1], and our principle-aware PRISM-scorer. For each design principle, we sample 200 examples (100 composed–perturbed pairs). Each model is queried to provide ratings for all five principles for every poster. We then compute the mean score for original and perturbed designs separately, and take their absolute difference (Δ) as a measure of sensitivity to that perturbation. A higher Δ indicates stronger discrimination between high- and low-quality designs for that principle. Refer to Supp. for prompt details.

4.4. Training of Localiser

We instruct-tune Qwen-2.5-VL (7B) using the localised annotations from the PRISM dataset, following the formulation described in Section 3.2.2. This experiment is conducted only for the *readability*, *contrast*, and *overlap* principles, as these can be spatially localised to specific text or visual elements within the layout. For each layout, the model predicts a set P of elements that violate the targeted

PRISM-scorer	Precision \uparrow	Recall \uparrow	F1-score \uparrow	AUC \uparrow
Coherence	0.7243	0.7279	0.7262	0.7983
Readability	0.7336	0.7323	0.7329	0.8064
Contrast	0.7552	0.7486	0.7519	0.8291
Alignment	0.7648	0.7591	0.7619	0.8321
Overlap	0.7175	0.7232	0.7203	0.7973

Table 1. **Performance of PRISM binary scorers.** Results on the held-out validation set showing Precision, Recall, F1, and AUC for each design principle.

principle, which is compared against the ground-truth set G from PRISM. True positives (TP) correspond to elements present in both P and G , false positives (FP) to elements in P but not in G , and false negatives (FN) to elements in G but missing from P . We compute Precision, Recall, and F1-score based on these counts, and Intersection-over-Union (IoU) using the sets to evaluate localization accuracy. To ensure consistency across models, we use the same system prompt for all baselines and instruct-tuned variants. Instruct-tuning is performed with one LoRA adapter [18] per principle, allowing independent specialization for *readability*, *contrast*, and *overlap*. We compare the instruct-tuned Qwen-2.5-VL model against three baselines: the pretrained Qwen-2.5-VL (base) model, a few-shot (3-shot) prompted version using the same instruction template, GPT-4o-mini, and GPT-4o queried with equivalent localised prompts. This comparison evaluates how PRISM perturbations enhance localisation.

5. Main Results

All analyses use the PRISM held-out validation split to assess responses to principle-specific degradations.

5.1. Principle Disentanglement: Human Evaluation

Figure 5 displays average human ratings across perturbation types and design principles. Diagonal scores are below the neutral midpoint of 4, indicating a strong preference for original designs over perturbed ones. For instance, coherence perturbations averaged 1.3 (coherence score), confirming their impact on thematic consistency. Off-diagonal ratings remained near 4, showing minimal effects on unrelated principles. We also observe a correlation between readability and contrast, and between alignment and overlap, which reflects the natural perceptual coupling among these visual attributes. These results show that PRISM perturbations introduce controlled, interpretable degradations that are perceptually distinct across design principles.

5.2. PRISM Scorer: Performance & Sensitivity

We first evaluate the PRISM binary scorers on the held-out validation set to assess their reliability across design principles. Qualitative examples illustrating these scorer responses are described in Figure 6(A). As shown in Table 1,

the scorers achieve moderate but consistent performance, with F1-scores ranging between 0.72 and 0.76 and AUC values between 0.80 and 0.82. While these results do not reflect perfect classification accuracy, the objective of the scorers is to achieve stable and interpretable sensitivity to principle-specific degradations, rather than absolute predictive performance. This sensitivity makes the scorers reliable quantitative indicators of principle-specific degradations.

Figure 7 visualizes model sensitivities across principles using radar plots. Each plot corresponds to one perturbation type and shows the absolute change in model ratings between composed and perturbed posters. The PRISM scorers exhibit the most principle-aligned behavior, responding strongly along the targeted dimension while remaining stable elsewhere. As each scorer is only trained on one specific principle, all other principles represent out-of-domain data. GPT-4o captures overall degradation trends but does not disentangle individual principles, and GPT-4o-mini and Qwen-2.5-VL show weaker, noisier responses. These findings indicate that the PRISM scorers capture fine-grained degradations and provide a basis for principle-aware evaluation of visual design quality. (Refer to Supp for additional OOD experiments).

5.3. PRISM: Localised Error Detection

Table 2 presents the localisation performance across the *readability*, *contrast*, and *overlap* principles. Instruct-tuning Qwen-2.5-VL with PRISM annotations substantially improves performance over the base, prompted, and GPT-based models, achieving the highest IoU and F1-scores across all principles. The instruct-tuned model achieves an average F1 score of 0.80 for *readability*, 0.72 for *contrast*, and 0.77 for *overlap*, demonstrating its ability to identify localised design violations accurately substantially outperforming GPT-4o which demonstrates moderate localisation ability but tends to produce less precise predictions. GPT-4o-mini and prompted Qwen-2.5-VL exhibit weaker and noisier trends. These results show that principle-specific fine-tuning with PRISM supervision leads to more interpretable and spatially grounded reasoning for detecting localised degradations. We also provide qualitative localiser outputs for these principles in Figure 6(B).

6. Demonstrative Editing Pipeline

To illustrate a possible utility of our framework, we implement a demonstrative editing pipeline for the coherence principle using a beam search-based strategy [44]. The pipeline uses feedback from the scorer and localisation modules to propose and evaluate principle-aware refinements, as introduced in Fig. 4(C). At each iteration, the localisation module identifies elements contributing to incoherence and helping generate candidate edits through prompt-based reasoning. Each candidate layout addresses

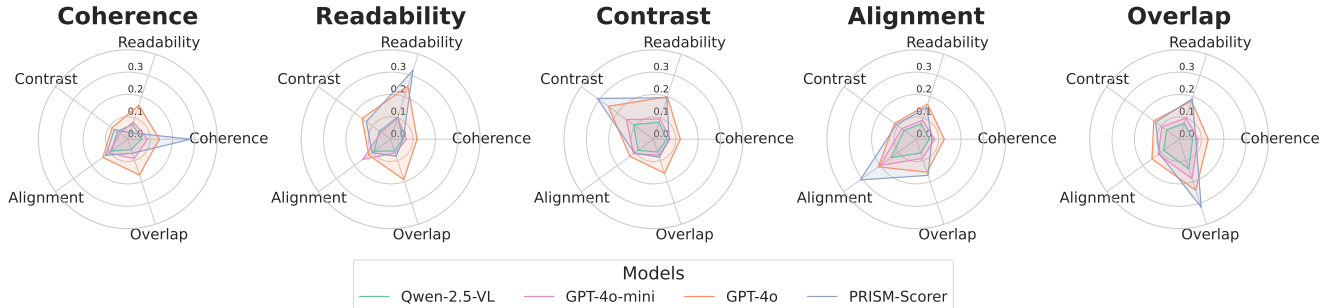


Figure 7. **Model sensitivity to PRISM perturbations.** Radar plots show average rating differences (Δ) between composed and perturbed posters across principles. Our PRISM scorer demonstrates the strongest and most principle-specific sensitivity.

Model	Readability				Contrast				Overlap			
	IoU \uparrow	Precision \uparrow	Recall \uparrow	F1-score \uparrow	IoU \uparrow	Precision \uparrow	Recall \uparrow	F1-score \uparrow	IoU \uparrow	Precision \uparrow	Recall \uparrow	F1-score \uparrow
Qwen _{Base}	0.3645	0.4091	0.5478	0.4312	0.2952	0.3314	0.4725	0.3804	0.3215	0.3411	0.4513	0.3770
Qwen _{Prompted}	0.3603	0.4303	0.4782	0.4147	0.3186	0.3519	0.4884	0.4049	0.3297	0.3520	0.4728	0.3984
GPT-4o-mini	0.4028	0.4387	0.6127	0.4718	0.3456	0.3924	0.5311	0.4478	0.3729	0.4012	0.5238	0.4526
GPT-4o	0.5532	0.6433	0.6228	0.6037	0.5164	0.5832	0.6215	0.6017	0.5417	0.6112	0.6354	0.6210
Qwen_{Expert} (Ours)	0.7833	0.8015	0.8132	0.7998	0.6761	0.7024	0.7452	0.7196	0.7328	0.7553	0.7914	0.7730
$\Delta_{\text{Ours-GPT-4o}}$	0.230\uparrow	0.158\uparrow	0.190\uparrow	0.196\uparrow	0.160\uparrow	0.119\uparrow	0.124\uparrow	0.118\uparrow	0.191\uparrow	0.144\uparrow	0.156\uparrow	0.152\uparrow

Table 2. **Localisation performance across principles.** Fine-tuning Qwen-2.5-VL with PRISM annotations leads to substantial improvements in IoU and F1 across readability, contrast, and overlap compared to base, prompted, GPT-4o-mini and GPT-4o.

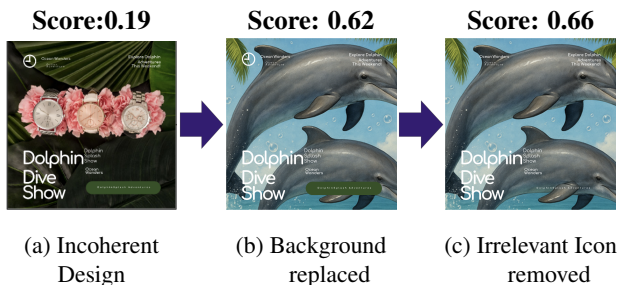


Figure 8. **Demonstrative coherence editing pipeline.** Starting from an incoherent design (a) from the PRISM dataset, the model iteratively refines the layout guided by the coherence scorer and localiser. The first edit (b) replaces the mismatched background with a theme-consistent image, while the second edit (c) removes the irrelevant watch icon. The coherence score increases from 0.19 \rightarrow 0.62 \rightarrow 0.66, reflecting improved thematic consistency.

one of the identified inconsistencies, yielding multiple alternative paths for editing. The coherence scorer re-evaluates these candidates, and the top- k candidates (with a beam size $k = 2$) achieving the highest coherence scores are retained for subsequent iterations. Unlike direct modifications from MLLMs, this pipeline preserves the underlying layout structure and retains the design’s editability throughout the refinement process. See Fig. 8 for an illustrative example, demonstrating how interpretable feedback enables targeted, principle-specific refinements in visual design.

7. Conclusions and Future Work

Conclusions. We introduce \blacklozenge PRISM, a principle-disentangled benchmark and a multi-scale evaluation framework for interpretable reasoning of visual designs. PRISM enables consistent comparison of disentanglement skills by introducing a principle-aware perturbed dataset. We combine lightweight scorers and instruction-tuned models for localising errors, which can be used to propose targeted edits. These components lay a foundation for future multimodal systems that can diagnose, explain, and improve visual layouts aligned with human designers.

Future Work. While PRISM offers controlled principle-specific perturbations, there are several opportunities for improvement. First, the assumption of strict principle independence is inexact (for instance, readability and contrast can influence one another). A possible avenue for future work is to model such interactions more explicitly. Second, current coherence assessments rely on prompt-based judgments that vary across generative models; greater robustness may be achieved through multiple annotators, more stable prompting strategies, or the incorporation of theme-level annotations. Third, the editing pipeline could be extended to support vector-level edits. Finally, the scorer and localiser can double as reward functions for RLHF- or DPO-style [36, 41] training, allowing models to learn scoring, localisation, and editing, holistically.

References

- [1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 6
- [2] Michael Bauerly and Yili Liu. Computational modeling and experimental investigation of effects of compositional elements on interface and design aesthetics. *Int. J. Hum. Comput. Stud.*, 64(8):670–682, 2006. 2
- [3] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm’s referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 3
- [4] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali: A jointly-scaled multilingual language-image model, 2023. 2
- [5] Sanjoy Chowdhury, Sayan Nag, Subhrajyoti Dasgupta, Jun Chen, Mohamed Elhoseiny, Ruohan Gao, and Dinesh Manocha. Meerkat: Audio-visual large language model for grounding in space and time. In *European Conference on Computer Vision*, pages 52–70. Springer, 2024. 3
- [6] Sanjoy Chowdhury, Mohamed Elmoghany, Yohan Abeysinghe, Junjie Fei, Sayan Nag, Salman Khan, Mohamed Elhoseiny, and Dinesh Manocha. Magnet: A multi-agent framework for finding audio-visual needles by reasoning over multi-video haystacks. *arXiv preprint arXiv:2506.07016*, 2025. 3
- [7] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining, 2025. 2
- [8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 2
- [9] Qi Dou, Xianjun Sam Zheng, Tongfang Sun, and Pheng-Ann Heng. Webthetics: Quantifying webpage aesthetics with deep learning. *Int. J. Hum. Comput. Stud.*, 124:56–66, 2019. 2
- [10] Peitong Duan, Chin-Yi Cheng, Gang Li, Bjoern Hartmann, and Yang Li. Uicrit: Enhancing automated design evaluation with a ui critique dataset. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pages 1–17, 2024. 2
- [11] Haraguchi et al. Can gpts evaluate graphic design based on design principles? In *SIGGRAPH Asia 2024*. 3
- [12] Felipe Gonzalez-Pizarro and Giuseppe Carenini. Neural multimodal topic modeling: A comprehensive evaluation. *ArXiv*, abs/2403.17308, 2024. 3
- [13] Sahil Goyal, Abhinav Mahajan, Swasti Mishra, Prateeksha Udhayan, Tripti Shukla, KJ Joseph, and Balaji Vasanth Srinivasan. Design-o-meter: Towards evaluating and refining graphic designs. 2025. 1, 2
- [14] Daichi Haraguchi, Naoto Inoue, Wataru Shimoda, Hayato Mitani, Seiichi Uchida, and Kota Yamaguchi. Can gpts evaluate graphic design based on design principles? In *SIGGRAPH Asia 2024 Technical Communications, SA 2024, Tokyo, Japan, December 3-6, 2024*, pages 5:1–5:4. ACM, 2024. 2
- [15] Steven J. Harrington, J. Fernando Naveda, Rhys Price Jones, Paul G. Roetling, and Nishant Thakkar. Aesthetic measures for automated document layout. In *Proceedings of the 2004 ACM Symposium on Document Engineering, Milwaukee, Wisconsin, USA, October 28-30, 2004*, pages 109–111. ACM, 2004. 2
- [16] Daichi Horita, Naoto Inoue, Kotaro Kikuchi, Kota Yamaguchi, and Kiyoharu Aizawa. Retrieval-augmented layout transformer for content-aware layout generation. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 67–76, 2023. 2
- [17] HsiaoYuan Hsu, Xiangteng He, Yuxin Peng, Hao Kong, and Qing Zhang. Posterlayout: A new benchmark and approach for content-aware visual-textual presentation layout, 2023. 2
- [18] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. 7
- [19] Danqing Huang, Jiaqi Guo, Shizhao Sun, Hanling Tian, Jieru Lin, Zheng Hu, Chin-Yew Lin, Jian-Guang Lou, and Dongmei Zhang. A survey for graphic design intelligence. *arXiv preprint arXiv:2309.01371*, 2023. 1, 3
- [20] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 2, 3, 6
- [21] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishal Shankar, Benjamin Recht, Simon Kumar, Ludwig Schmidt, and Alessandro Ross. Openclip, 2021. 2
- [22] Peidong Jia, Chenxuan Li, Zeyu Liu, Yichao Shen, Xingru Chen, Yuhui Yuan, Yinglin Zheng, Dong Chen, Ji Li, Xiaodong Xie, Shanghang Zhang, and Baining Guo. COLE: A hierarchical generation framework for graphic design. *CoRR*, abs/2311.16974, 2023. 2
- [23] Wenyuan Kong, Zhaoyun Jiang, Shizhao Sun, Zhuoning Guo, Weiwei Cui, Ting Liu, Jianguang Lou, and Dongmei Zhang. Aesthetics++: Refining graphic designs by exploring design principles and human preference. *IEEE Trans. Vis. Comput. Graph.*, 29(6):3093–3104, 2023. 2
- [24] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023. 3

- [25] Jianan Li, Jimei Yang, Jianming Zhang, Chang Liu, Christina Wang, and Tingfa Xu. Attribute-conditioned layout gan for automatic graphic design. *IEEE Transactions on Visualization and Computer Graphics*, 27:4039–4048, 2020. 2
- [26] William Lidwell, Kritina Holden, and Jill Butler. *Universal Principles of Design*. Rockport Publishers, second edition edition, 2010. 3
- [27] Chenguo Lin, Yuchen Lin, Panwang Pan, Xuanyang Zhang, and Yadong Mu. Instructlayout: Instruction-driven 2d and 3d layout synthesis with semantic graph prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47: 11040–11050, 2024. 2
- [28] Simon Lok and Steven K. Feiner. A survey of automated layout techniques for information presentations. 2005. 2
- [29] Shehan Munasinghe, Rusiru Thushara, Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, Mubarak Shah, and Fahad Khan. Pg-video-llava: Pixel grounding large video-language models. *arXiv preprint arXiv:2311.13435*, 2023. 3
- [30] Sayan Nag, KJ Joseph, Koustava Goswami, Vlad I Morariu, and Balaji Vasan Srinivasan. Agentic design review system. *arXiv preprint arXiv:2508.10745*, 2025. 1, 2
- [31] David Chek Ling Ngo, Lian Seng Teo, and John G Byrne. A mathematical theory of interface aesthetics. In *Visual mathematics*. Mathematical Institute SASA, 2000. 2
- [32] Peter O’Donovan, Aseem Agarwala, and Aaron Hertzmann. Designscape: Design with interactive layout suggestions. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015. 2
- [33] OpenAI. Gpt-4o mini. <https://openai.com>, 2024. Accessed: 2025-02-13. 2
- [34] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 2
- [35] Mayu Otani, Naoto Inoue, Kotaro Kikuchi, and Riku Togashi. Ltsim: Layout transportation-based similarity measure for evaluating layout generation. *arXiv preprint arXiv:2407.12356*, 2024. 1
- [36] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. 8
- [37] Peter O’Donovan, Aseem Agarwala, and Aaron Hertzmann. Learning layouts for single-page graphic designs. *IEEE transactions on visualization and computer graphics*, 20(8): 1200–1213, 2014. 2
- [38] Sohan Patnaik, Rishabh Jain, Balaji Krishnamurthy, and Mausoom Sarkar. Aestheticq: Enhancing graphic layout design via aesthetic-aware preference alignment of multimodal large language models, 2025. 1
- [39] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 3
- [40] Shraman Pramanick, Guangxing Han, Rui Hou, Sayan Nag, Ser-Nam Lim, Nicolas Ballas, Qifan Wang, Rama Chellappa, and Amjad Almahairi. Jack of all tasks master of many: Designing general-purpose coarse-to-fine vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14076–14088, 2024. 3
- [41] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. 8
- [42] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Erix Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. *arXiv preprint arXiv:2311.03356*, 2023. 3
- [43] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2015. 2
- [44] Chufan Shi, Haoran Yang, Deng Cai, Zhisong Zhang, Yifan Wang, Yujiu Yang, and Wai Lam. A thorough examination of decoding methods in the era of llms. *arXiv preprint arXiv:2402.06925*, 2024. 2, 7
- [45] Yong Shi, Mengyuan Shang, and Zhiquan Qi. Intelligent layout generation based on deep generative models: A comprehensive survey. *Inf. Fusion*, 100:101940, 2023. 2
- [46] Kihoon Son, DaEun Choi, Tae Soo Kim, and Juho Kim. Demystifying tacit knowledge in graphic design: Characteristics, instances, approaches, and guidelines. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2024. 1, 3
- [47] Manan Suri, Puneet Mathur, Franck Dernoncourt, R. Jain, Vlad I Morariu, Ramit Sawhney, Preslav Nakov, and Dinesh Manocha. Docedit-v2: Document structure editing via multimodal llm grounding. In *Conference on Empirical Methods in Natural Language Processing*, 2024. 2
- [48] Sou Tabata, Hiroki Yoshihara, Haruka Maeda, and Kei Yokoyama. Automatic layout generation for graphical design magazines. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference, SIGGRAPH 2019, Los Angeles, CA, USA, July 28 - August 1, 2019, Posters*, pages 9:1–9:2. ACM, 2019. 2
- [49] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 2, 6

- [50] Weiyun Wang, Min Shi, Qingyun Li, Wenhai Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, et al. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. *arXiv preprint arXiv:2308.01907*, 2023. 3
- [51] Shuning Xue, Bu Jin, Jie Jiang, Longteng Guo, and Jing Liu. A hybrid local-global neural network for visual classification using raw eeg signals. *Scientific Reports*, 14, 2024. 3
- [52] Kota Yamaguchi. Canvasvae: Learning to generate vector graphic documents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5481–5489, 2021. 2, 3, 6
- [53] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023. 3
- [54] Mathieu Zen and Jean Vanderdonckt. Towards an evaluation of graphical user interfaces aesthetics based on metrics. In *IEEE 8th International Conference on Research Challenges in Information Science, RCIS 2014, Marrakech, Morocco, May 28-30, 2014*, pages 1–12. IEEE, 2014. 2
- [55] Hui Zhang, Dexiang Hong, Maoke Yang, Yutao Cheng, Zhao Zhang, Jie Shao, Xinglong Wu, Zuxuan Wu, and Yu-Gang Jiang. Creatidesign: A unified multi-conditional diffusion transformer for creative graphic design, 2025. 6, 2
- [56] Jiahao Zhang, Ryota Yoshihashi, Shunsuke Kitada, Atsuki Osanai, and Yuta Nakashima. Vascar: Content-aware layout generation via visual-aware self-correction. *arXiv preprint arXiv:2412.04237*, 2024. 1, 2
- [57] Nanxuan Zhao, Ying Cao, and Rynson W. H. Lau. What characterizes personalities of graphic designs? *ACM Trans. Graph.*, 37(4):116, 2018. 2
- [58] Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. Bubogpt: Enabling visual grounding in multi-modal llms. *arXiv preprint arXiv:2307.08581*, 2023. 3
- [59] Xingxing Zou, Wen Zhang, and Nanxuan Zhao. From fragment to one piece: A review on ai-driven graphic design. *Journal of Imaging*, 11(9):289, 2025. 1

Through the PRISM: Principle-Aware, Interpretable, and Multi-Scale Evaluation of Visual Designs

Supplementary Material

The supplementary is organised as follows:

- A PRISM Perturbations
- B Prompting Setup for Model Sensitivity
- C Backbone Ablation
- D Localised Error Detection Details
- E Scorer Out-of-Domain Generalization
- F Additional Examples from Editing Pipeline

A. PRISM Perturbations

In this section we provide additional details about how each perturbation in PRISM is constructed. The main paper introduces the framework and motivation in Section 3.1. Below, we describe the procedures used to isolate each design principle while keeping all other aspects of the poster unchanged so that each variant reflects a targeted violation of only one dimension. For the perturbations that provide localized supervision (readability, contrast, and overlap), we also record the elements involved so they can be used during instruction-tuning (Section 3.2.2).

A.1. Coherence

For coherence perturbation, we reassign each poster a new semantic theme sampled from a different category. We apply either a text-based or element-based modification selected at random. In the text perturbation, all textual content is rewritten using a few-shot LLM prompt so that the new text reflects the target theme while maintaining similar length. Only the text in the metadata is modified, thus preserving attributes such as font style, boldness, alignment, spacing, and color. For element modification, we identify visually prominent components using a vision-language query over the composed layout and check their relevance to the original theme using a lightweight LLM-based relevance classifier. We remove or replace elements determined to be theme-specific, and important background regions are regenerated using an image editor conditioned on the new theme while maintaining the color family. The new design preserves structure but conveys a different semantic theme.

A.2. Readability

Readability perturbation reduces the legibility of text while keeping the overall layout intact. For each poster, we randomly select a subset of text elements and apply one or more readability changes. These include shrinking text that is unusually large, adjusting the line height of multi-line text

blocks so that spacing becomes compressed, and shifting of text color toward the dominant background color behind it. Each poster receives only a subset of these modifications, and the perturbations applied to each text block are recorded so that the dataset includes explicit supervision about how readability was degraded.

A.3. Contrast

Contrast perturbation weakens the visual separation between foreground and background while preserving the original structural layout. We randomly select a subset of text, icons, or decorative elements and apply one or more transformations. Text contrast perturbation is similar to readability perturbation with contrast. We edit icons and other graphical components using a diffusion-based tool that adjusts their color distributions so that they blend more closely with surrounding areas. For each modification, we record the specific foreground-background pairs where contrast was intentionally reduced.

A.4. Alignment

Alignment perturbation introduces positional shifts that disrupt the structural consistency of the design. For each poster, we randomly select a small set of elements and shift them horizontally or vertically so that they break expected alignment patterns such as column structure or centered grouping. When a text block has an element, that is not the overall background, just below it, we move both together so that the perturbation affects alignment without introducing unintended overlap or contrast issues. These modifications do not map cleanly onto specific localised pairs but instead makes changes to the poster on a global level.

A.5. Overlap

Overlap perturbation introduces unintended occlusions between non-text elements while keeping all other aspects of the layout unchanged. We randomly select a small number of icons, shapes, or decorative components and position them so that they partially cover other elements while maintaining the overall alignment. This produced cases such as icon overlapping with icon or shape overlapping with another object, hence reducing visual separability. For each overlap introduced, we record the specific element pairs involved for explicit supervision.

SYSTEM

You are an expert evaluator for visual design quality. Given one design image (for example a poster, flyer, or layout), assess the following principles of design quality. For each principle, assign a continuous score between 0 and 1 (0 = very poor, 1 = excellent). A score below 0.5 indicates that the design needs improvement.

- 1.Coherence (0-1) — How well does the overall layout align with the design's theme, purpose, and visual hierarchy? Lower scores indicate the layout feels disjointed or confusing.
- 2.Readability (0-1) — How easy is it to read and process textual information? Lower scores suggest text is hard to read due to font choice, size, color, or placement.
- 3.Contrast (0-1) — How well do visual elements stand out through tonal, color, or spatial contrast? Lower scores imply elements blend together or lack emphasis.
- 4.Alignment (0-1) — How consistent and logical is the placement of elements? Lower scores reflect misaligned or haphazardly placed items.
- 5.Overlap (0-1) — How cleanly are visual layers and elements arranged without unwanted obstruction? Lower scores indicate important elements are obscured or cluttered.

Return the response in JSON format:
 {"coherence_score": float, "readability_score": float, "contrast_score": float, "alignment_score": float, "overlap_score": float}

USER

Text: Provide scores without any explanation based on the rubric. Output JSON only.
Image: <Design-input>

ASSISTANT

```
{
  "coherence_score": 0.7,
  "readability_score": 0.2,
  "contrast_score": 0.4,
  "alignment_score": 0.8,
  "overlap_score": 0.9
}
```

Figure 9. Prompts for evaluating model sensitivity to PRISM Perturbations. For results, see Fig 3.

B. Prompting Setup for Model Sensitivity

To evaluate model sensitivity to PRISM perturbations (Section 4.3), we query each model under a consistent rubric-based prompting setup. The system prompt defines the scoring criteria for all five design principles and the overall score, each on a continuous scale between 0 and 1. The user prompt requests that the model return a JSON response without any explanation. For every composed or perturbed poster, the corresponding image is supplied directly to the model using its multimodal interface. For reproducibility, we include the exact prompts used for GPT-4o, GPT-4o-mini, and Qwen-2.5-VL in Figure 9. We also include average scores and sensitivity (Δ) from each model across design principles in Table 4.

C. Backbone Ablation

Table 3 compares several visual backbones. This ablation evaluates whether architectural choice or multimodal pre-

	Backbone	Precision \uparrow	Recall \uparrow	F1-score \uparrow	AUC \uparrow
Frozen	ViT-B/16	0.528	0.522	0.524	0.554
	DINOv2	0.539	0.531	0.535	0.563
	OpenCLIP ViT-B/16	0.551	0.544	0.547	0.577
	SigLIP-v2	0.556	0.548	0.552	0.584
Fine-tuned	OpenCLIP ViT-B/16	0.701	0.693	0.697	0.781
	SigLIP-v2 (Ours)	0.7391	0.7382	0.7386	0.8126

Table 3. **Backbone comparison for PRISM-scorer.** Untrained backbones show limited design-sensitivity, while pretrained contrastive models (CLIP, SigLIP-v2) demonstrate strong performance. **SigLIP-v2 (Ours)** achieves the best overall scores across all metrics.

training affects a model’s ability to detect principle-specific design degradations. In the frozen setting, each backbone is initialized with its standard pretrained weights: ViT-B/16 [8] trained on ImageNet [43], SigLIP-v2 [49] trained on WebLI [4], and the self-supervised DINOv2 model [34], but we freeze the entire backbone and train only an identical lightweight two-layer binary classifier on top for all models, ensuring that differences arise solely from the backbone representations. We then report averaged PRISM-scorer performance across all five principles. These frozen models provide limited but non-trivial sensitivity to design perturbations, with vision-only backbones (ViT-B/16, DINOv2) performing the weakest and frozen vision–language architectures (OpenCLIP [21], SigLIP-v2) offering slightly stronger baselines due to their multimodal structure. In the fine-tuned setting, we update all backbone weights using the principle-aware fine-tuning procedure described in Section 3.2.1, leveraging paired image–caption data from Crello [52] and CreatiDesign [55]. This multimodal contrastive supervision leads to substantial improvements: fine-tuned OpenCLIP ViT-B/16 significantly outperforms all frozen variants, and SigLIP-v2 (Ours) achieves the strongest performance across precision, recall, F1, and AUC. These findings underscore the importance of principle-aware multimodal fine-tuning and motivate our use of SigLIP-v2 as the backbone for PRISM-scorer.

We also compare the simple linear head on top of this design-aware architecture with other variants. Averaged across principles, the test F1 scores are: linear head (73.6%), two-layer MLP (71.4%), and multi-head attention (72.5%). These results show that a simpler linear layer is sufficient to achieve the required performance gains.

We compare a unified scorer–localizer model, evaluated by training Qwen-2.5-VL end-to-end. The unified model achieves an average F1 score of 65% as a scorer and mean IoU scores of 60% across principles. The unified model performs worse than our expert PRISM scorer and localizer models. This explains the need for experts instead of one unified model trained for multiple tasks.

Model	Coherence			Readability			Contrast			Alignment			Overlap		
	Org	Perturb	Δ	Org	Perturb	Δ	Org	Perturb	Δ	Org	Perturb	Δ	Org	Perturb	Δ
Qwen-2.5-VL	0.739	0.678	0.061	0.807	0.736	0.071	0.742	0.684	0.058	0.780	0.690	0.090	0.718	0.669	0.049
GPT-4o-mini	0.751	0.662	0.088	0.678	0.597	0.081	0.635	0.590	0.045	0.726	0.622	0.104	0.782	0.691	0.091
GPT-4o	0.824	0.683	0.141	0.693	0.533	0.160	0.759	0.671	0.087	0.827	0.689	0.138	0.787	0.619	0.168
PRISM-Scorer	0.658	0.377	0.281	0.633	0.602	0.031	0.712	0.638	0.074	0.743	0.621	0.122	0.655	0.589	0.066

(a) Coherence Results.

Model	Coherence			Readability			Contrast			Alignment			Overlap		
	Org	Perturb	Δ	Org	Perturb	Δ	Org	Perturb	Δ	Org	Perturb	Δ	Org	Perturb	Δ
Qwen-2.5-VL	0.744	0.690	0.054	0.830	0.725	0.105	0.751	0.689	0.062	0.795	0.703	0.092	0.736	0.681	0.055
GPT-4o-mini	0.751	0.680	0.071	0.675	0.575	0.100	0.633	0.580	0.053	0.734	0.580	0.154	0.784	0.720	0.064
GPT-4o	0.822	0.704	0.118	0.709	0.459	0.250	0.784	0.626	0.158	0.826	0.705	0.121	0.795	0.604	0.190
PRISM-Scorer	0.599	0.543	0.056	0.664	0.340	0.324	0.712	0.578	0.134	0.643	0.541	0.102	0.725	0.645	0.080

(b) Readability Results.

Model	Coherence			Readability			Contrast			Alignment			Overlap		
	Org	Perturb	Δ	Org	Perturb	Δ	Org	Perturb	Δ	Org	Perturb	Δ	Org	Perturb	Δ
Qwen-2.5-VL	0.751	0.691	0.06	0.810	0.730	0.08	0.720	0.610	0.11	0.720	0.635	0.085	0.720	0.660	0.06
GPT-4o-mini	0.752	0.677	0.075	0.630	0.535	0.095	0.610	0.460	0.15	0.735	0.625	0.11	0.810	0.735	0.075
GPT-4o	0.810	0.690	0.12	0.720	0.520	0.2	0.790	0.540	0.25	0.790	0.660	0.13	0.750	0.590	0.16
PRISM-Scorer	0.620	0.550	0.07	0.610	0.415	0.195	0.710	0.400	0.31	0.670	0.565	0.105	0.710	0.625	0.085

(c) Contrast Results.

Model	Coherence			Readability			Contrast			Alignment			Overlap		
	Org	Perturb	Δ	Org	Perturb	Δ	Org	Perturb	Δ	Org	Perturb	Δ	Org	Perturb	Δ
Qwen-2.5-VL	0.748	0.690	0.058	0.804	0.732	0.072	0.734	0.667	0.067	0.751	0.611	0.14	0.717	0.652	0.065
GPT-4o-mini	0.720	0.638	0.082	0.652	0.562	0.09	0.641	0.561	0.08	0.742	0.542	0.2	0.827	0.737	0.09
GPT-4o	0.815	0.690	0.125	0.731	0.566	0.165	0.781	0.661	0.12	0.803	0.593	0.21	0.731	0.576	0.155
PRISM-Scorer	0.652	0.582	0.07	0.640	0.490	0.15	0.721	0.611	0.11	0.692	0.382	0.31	0.728	0.558	0.17

(d) Alignment Results.

Model	Coherence			Readability			Contrast			Alignment			Overlap		
	Org	Perturb	Δ	Org	Perturb	Δ	Org	Perturb	Δ	Org	Perturb	Δ	Org	Perturb	Δ
Qwen-2.5-VL	0.732	0.672	0.06	0.821	0.746	0.075	0.730	0.660	0.07	0.756	0.671	0.085	0.750	0.610	0.14
GPT-4o-mini	0.762	0.677	0.085	0.625	0.525	0.1	0.658	0.563	0.095	0.752	0.637	0.115	0.783	0.598	0.185
GPT-4o	0.809	0.679	0.13	0.713	0.533	0.18	0.774	0.634	0.14	0.832	0.682	0.15	0.784	0.544	0.24
PRISM-Scorer	0.630	0.555	0.075	0.656	0.466	0.19	0.741	0.611	0.13	0.643	0.533	0.11	0.732	0.412	0.32

(e) Overlap Results.

Table 4. **Average Model Sensitivity Scores.** The table shows the scores from different models on original vs perturbed layouts across design principles, Δ denotes the difference between the averages showing sensitivity. While GPT-4o shows sensitivity, PRISM-Scorer is able to disentangle the principles better than any other model.

D. Localised Error Detection Details

We get the data required for instruction-tuning as described in Section A. To train the model on localised supervision from readability, contrast and overlap perturbations, we use the prompts as shown in Figure 10. For readability, the corresponding image input is simply the design, whereas for contrast and overlap, we annotate the design using the metadata to include IDs for each non-textual element. For readability, the model is asked to choose from a list of texts present in the design, while for contrast and overlap, we provide pairs of elements to choose from. The list only contains pairs which have some overlapping components in order to contain the combinations to a reasonable number. We use the same prompt for evaluating and comparing different models on the held-out set. We use the training and

evaluating methods as mentioned in Section 4.4 and 5.3, respectively, for the results refer to Figure 2.

E. Scorer Out-of-domain generalization

As observed in Figure 7, the PRISM scorer is able to disentangle design principles. We evaluate the coherence scorer with all other perturbations that it has not seen during training, giving us insights into the scorer’s generalization performance. While the scorer can generalize to all types of PRISM perturbations, we test the potential of the scorer on a different perturbation style from the GDE dataset [11]. This dataset only contains overlap and alignment perturbations. We evaluate our scorers using human annotations from the GDE dataset as ground truth. The PRISM scorer for overlap and alignment achieves F1 scores of 72% and 71%,

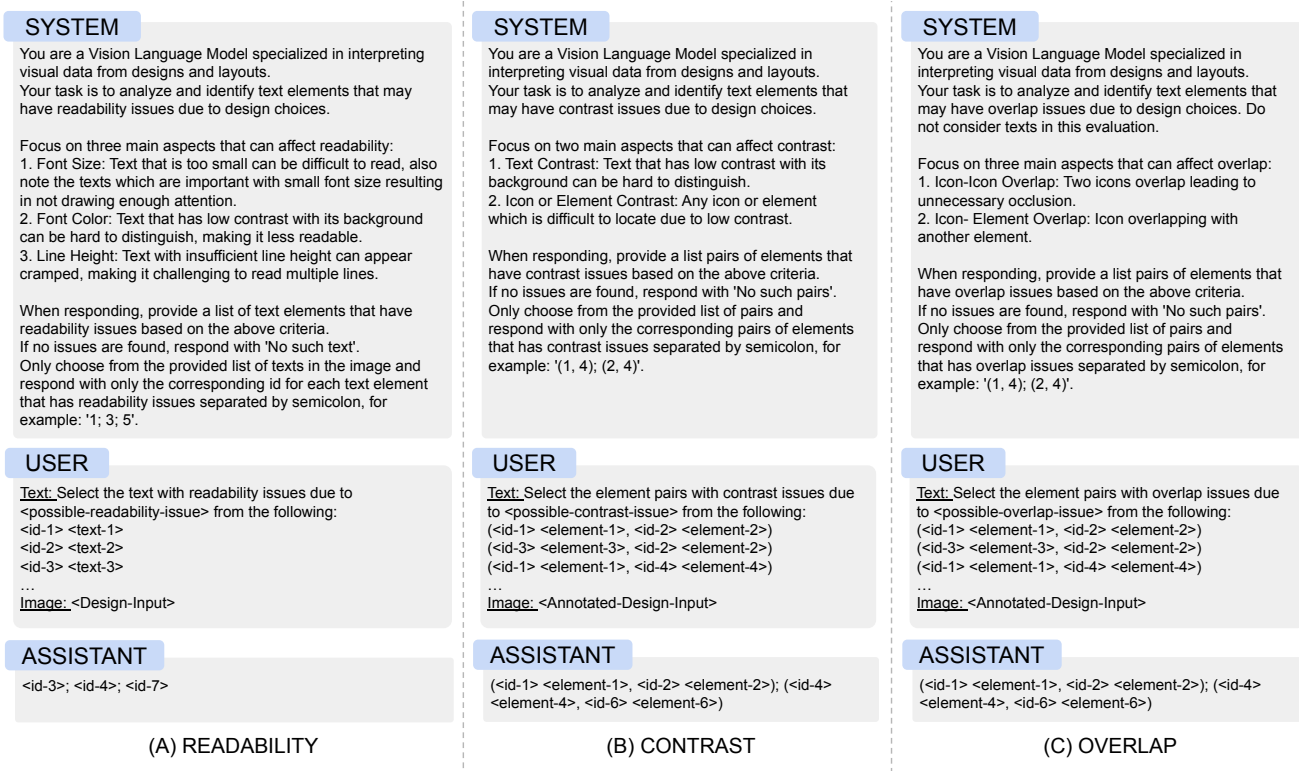


Figure 10. **Prompts for Localising Error Detection.** The figure shows the prompts used for instruct-tuning and evaluating models for local principles (Readability (A), Contrast (B), Overlap (C)). Results comparing different models using this prompt can be seen in Table 2.

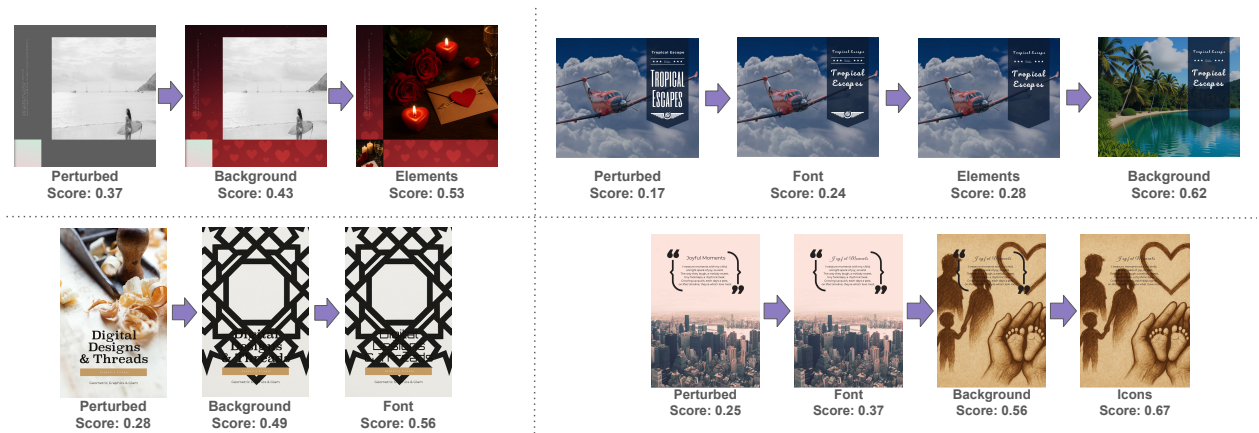


Figure 11. **Coherence-based editing examples.** The grid shows perturbed posters and the editing path (See Section 5.4). Since the pipeline focuses only on coherence, edits target the background, elements, and font for thematic consistency, while other principles remain unchanged. Scores reflect the scorer’s coherence predictions at each step.

respectively, showing generalization to other types of perturbations as well.

F. Additional Examples from Editing Pipeline

As shown in Figure 11, we present additional examples from our demonstrative editing pipeline, which operates

exclusively using the coherence module. Candidate edits are proposed by the coherence localizer and ranked at each step by the coherence scorer, resulting in refinements that focus solely on restoring thematic and stylistic consistency (See Section 5.4). These examples only look at coherence-driven changes, while keeping the other princi-

ples untouched. A natural next step is to extend the editing pipeline to all design principles and explore ways to combine these principle-specific refinements. A unified system that can coordinate edits across principles would enable more comprehensive and well-rounded layout improvement.