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

Supplementary Material

The supplementary is organised as follows:

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-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.

	Backbone	Precision ↑	Recall ↑	F1-score ↑	AUC ↑
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.

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 %,

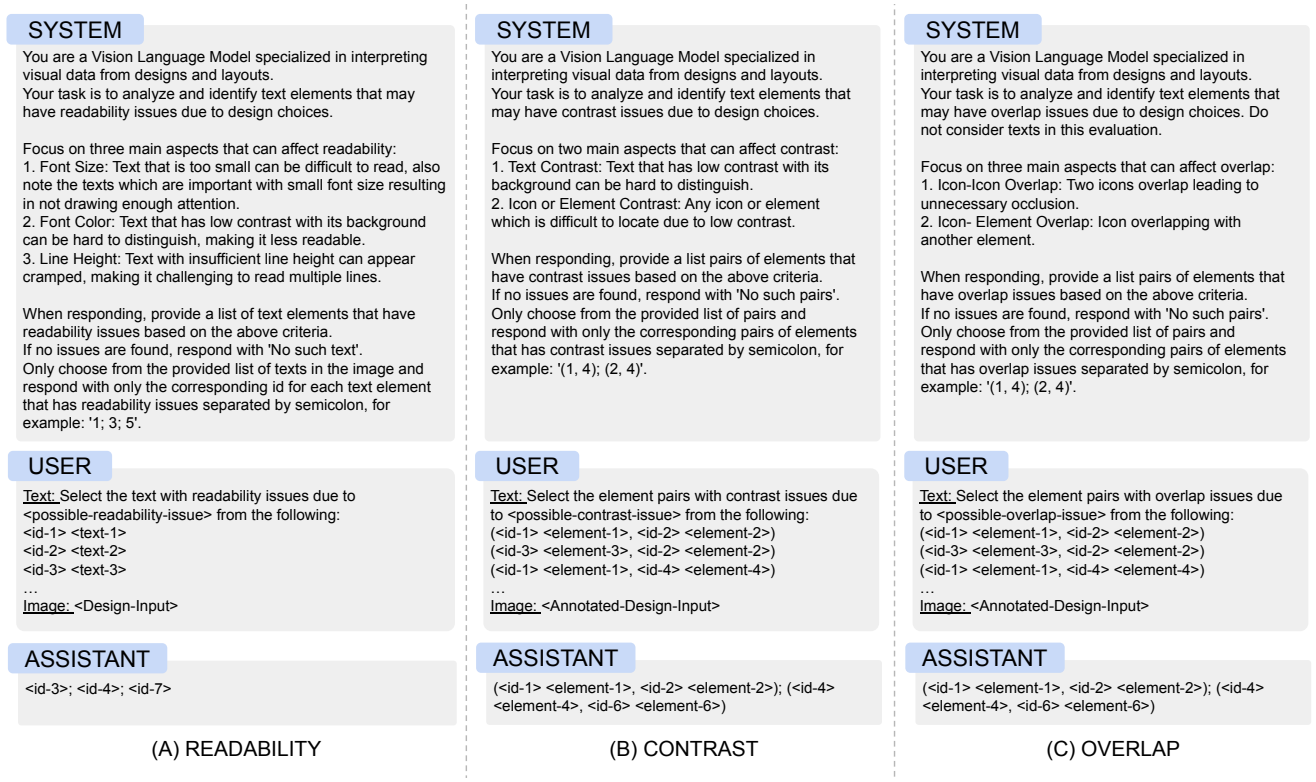

Figure 10. **Prompts for Localising Error Detection.** The figure shows the prompts used for instruct-tuning and evaluating models for local principles (Readability (A), Contrast (B), Overlap (C)). Results comparing different models using this prompt can be seen in Table 2.

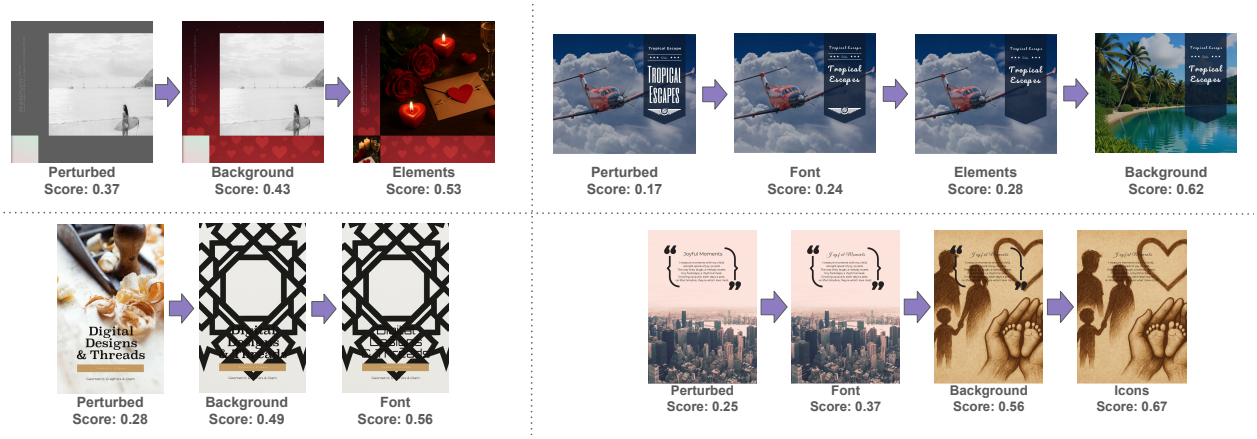

Figure 11. **Coherence-based editing examples.** The grid shows perturbed posters and the editing path (See Section 5.4). Since the pipeline focuses only on coherence, edits target the background, elements, and font for thematic consistency, while other principles remain unchanged. Scores reflect the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