Agentic Design Review System

Sayan Nag, Joseph K J, Koustava Goswami, Vlad I Morariu, Balaji Vasan Srinivasan

Adobe Research

Abstract

Evaluating a graphic design involves assessing it from multiple facets like alignment, composition, aesthetics and color choices. Holistic evaluation would involve aggregating feedback from individual expert reviewers. Towards this, we propose an Agentic Design Review System (Agentic-DRS), where multiple agents collaboratively analyze a design, orchestrated by a meta-agent. A novel in-context exemplar selection approach based on graph matching and a unique prompt expansion method plays central role towards making each agent design aware. In order to evaluate this framework, we propose DRS-BENCH . Thorough experimental evaluation against state-of-the-art baselines adapted to the problem setup, backed by critical ablations, demonstrates efficacy of Agentic-DRS in evaluating designs and generating actionable feedback.

1 Introduction

Graphic designs like flyers, posters, invitation-cards, etc., are harmonious compositions of images, text, shapes and their colors, nicely laid-out aesthetically, to convey the meaning intended by their designer. They have become ubiquitous in our daily lives from the brochure of a new car to the birthday invitation of a toddler. With the availability of do-it-yourself design tools, amateurs and novice designers are empowered to create professional designs. With the proliferation in the use of such designs in social media platforms, these tools are getting increasingly popular. Novice designers lack deep understanding of design principles like balance, emphasis, unity, white-space usage and so on, which would have a profound impact of their final generation. A tool that can *automatically analyze a design and provide actionable feedback* would be of immense value for such designers.

With the recent advancements in Multi-modal LLMs (Liu et al. 2023; Zhang et al. 2024) and Diffusion Models (Rombach et al. 2022), researchers have introduced novel approaches (Jia et al. 2023b; Inoue et al. 2024) for generating graphic designs from textual prompts. This is a significant advancement from the earlier efforts (Inoue et al. 2023b; Levi et al. 2023; Luo et al. 2024; Lin et al. 2023) which generate just the layout (positioning information of elements), to generating the entire design, with the content filled in. As these technologies mature, they would truly enable Human-AI co-creation for design generation. A yard-stick to measure the progress of design generation methods would be a design-evaluator that *introspects a design across multiple dimensions* such as typographic quality, color consistency and semantic coherence.

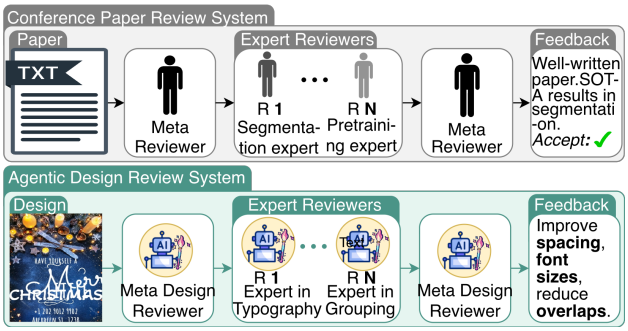

Figure 1: Evaluating a graphic design involves assessing it across multiple dimensions like visual coherence, semantic grouping, typographic clarity and so on. Inspired by peer-review system in conferences, we propose to build the *first Agentic framework for design evaluation and feedback generation*. We propose a novel approach to infuse design knowledge into agents, and allow them to collaboratively review and access the input design in Sec 3.

Judging whether a design is good or bad is hard and subjective as characteristics of designs are inherently *tacit*[1] (Son et al. 2024). Tacit principles like color pairing for a specific demography, is hard to objectively define and quantify making design evaluation extremely challenging. Heuristic approaches (O'Donovan, Agarwala, and Hertzmann 2014; Ngo, Teo, and Byrne 2000) try to mathematically quantify design characteristics like alignment, overlap and whitespace to give a score, but fail to see the overall global harmony of the designs. Learning based approaches (Zhao, Cao, and Lau 2018a; Sawada et al. 2024) treats the design evaluation as either a regression or classification task. Presence of a diverse enough dataset is a prerequisite for its performance. Recently, evaluations form GPT-4o (Hurst et al. 2024) has found to correlate well with human judgment of graphic de-

[1]Dictionary meaning: understood/implied without being stated.

signs (Haraguchi et al. 2024). Building on this trend, we propose a holistic evaluation suite to access the quality of graphic designs, and generate actionable feedback for the designers.

Due to the huge amount of training data and the learning paradigm, Multimodal LLMs (MLLMs) such as GPT-4o (Hurst et al. 2024) possess novice level awareness of the characteristics of good graphic design. Our first contribution is to enhance the design awareness of these models using *a novel graph matching based in-context exemplar selection approach* (Sec 3) and *a structured description based prompt expansion strategy* (Sec 3). A single instance of such a design-aware MLLM would not be able to assess a design across the variety of dimensions on which graphic designs should be evaluated. Toward this end, we propose the *first Agentic framework for design evaluation and feedback generation*. Each agent will be specializing in a specific aspect of evaluation like color harmony, typographic quality, alignment consistency and so on. Given an input design, a meta-agent spawns off these agents as necessary (could even dynamically control what aspect the agent should evaluate a design on, referred to as dynamic agents in Sec 3), and collates the independent feedback after the review process to generate scores and actionable feedback.

We evaluate our approach by introducing DRS-BENCH , a holistic benchmark suite containing 15 design attribute definitions, 4 datasets, new evaluation metrics and strong baselines. Our experimental analysis showcases the effectiveness of using design-aware MLLMs in an Agentic framework for design evaluation and feedback generation.

Our key contributions are summarized below:

- We introduce the first Agentic evaluation framework that can score designs and generate actionable feedback
- We enhance the design awareness of Multi-modal LLMs with a novel graph matching based exemplar selection approach and structured description based prompt expansion
- We introduce DRS-BENCH , a holistic framework for assessing design evaluation quality and feedback generation
- Through rigorous experimentation, we bring out the efficacy of our proposed approach, clearly out-performing the state-of-the-art baselines adapted to the task.

2 Related Works

Design Evaluation. Design evaluation has been explored across domains and development stages. Early heuristic methods (O'Donovan, Agarwala, and Hertzmann 2014; Zen and Vanderdonckt 2014) failed to capture semantic attributes. ML models (Dou et al. 2019) improved this but lacked comparative scoring, prompting siamese-based approaches (Zhao, Cao, and Lau 2018b; Goyal et al. 2025). Others focused on aesthetics (Kong et al. 2023) or used layout perturbation for scoring (Tabata et al. 2019). Recent MLLM-based methods target design generation (Jia et al. 2023a) or assess limited attributes (Jia et al. 2023a; Haraguchi et al. 2024), but lack comprehensive, actionable feedback - a gap we address.

MLLM Agents. Agentic workflows with large models are increasingly used for reasoning tasks. While early visual agents were task-specific (Shridhar et al. 2020), recent MLLMs support broader multi-modal agentic workflows (Xie et al. 2024). These models enable agents to handle diverse tasks, including design (Si et al. 2024; Laurençon, Tronchon, and Sanh 2024), and audio understanding (Zhang et al. 2023), using structured planning and reasoning. However, to the best of our knowledge, no prior work has explored a multi-agent, multi-modal workflow for Design Evaluation. We take the first step in this direction.

3 Methodology

Evaluating designs and providing actionable feedback is inherently complex, requiring nuanced understanding of aesthetics, functionality, user expectations, and design intricacies. Haraguchi et al. (2024) demonstrates how MLLMs like GPT-4o can evaluate designs in alignment with human ratings, marking a key step toward systematic design evaluation. Building on this, we enhance MLLM design capabilities in a *training-free* manner through two innovations: a novel graph-based in-context design selection method (Sec 3) and visually grounded structured design descriptions (Sec 3). Together, these enable dynamic retrieval of semantically and structurally relevant designs, improving contextual understanding, evaluation accuracy and robustness.

Design evaluation should consider multiple aspects such as visual appeal, semantic coherence, typographic quality, etc. To generate holistic feedback, we draw inspiration from human peer-review processes, where expert reviewers with diverse specializations assess the same work from different perspectives, and meta-reviewers aggregate their opinions. This motivates our proposal of an Agentic Design Review System (Fig. 1) - to our knowledge, the first formalized collaborative review framework for graphic design evaluation (Sec. 3). By mirroring expert-driven decision-making principles, AGENTIC-DRS enhances evaluation robustness and fosters a more explainable, structured critique process. It enables nuanced feedback that balances subjective creativity with objective principles, offering designers context-aware, multi-faceted insights for meaningful improvements.

Graph-based Design Exemplar Selection

Conventional in-context example selection often relies on global CLIP feature similarity, which can miss finer design aspects like spatial composition, alignment, grouping, reading order, and contextual interplay. Our **GRA**ph-based **D**esign exemplar selection method (GRAD) addresses this by encoding *semantic*, *spatial*, and *structural* relationships through localized graph representations of graphic designs. GRAD offers key advantages: *(i)* Preservation of structural and semantic relationships - unlike global feature matching, our graph captures relative positions and proximity of design elements; *(ii)* Adaptability to open-world designs - by combining semantic and spatial embeddings, GRAD generalizes across diverse design styles. For structured comparison and retrieval, we represent the Query design $\mathcal{D}^{Q}$ and In-Context library designs $\mathcal{D}^{\text{IC}} \in \mathcal{D}$ as graphs $\mathcal{G}_{\mathcal{D}^j} = (\mathbb{V}^j, \mathbb{E}^j)$, where $\mathbb{V}$ and $\mathbb{E}$ denote vertices and edges, and $\mathcal{D}$ is the superset of all designs. Graph creation is done following two strategies:

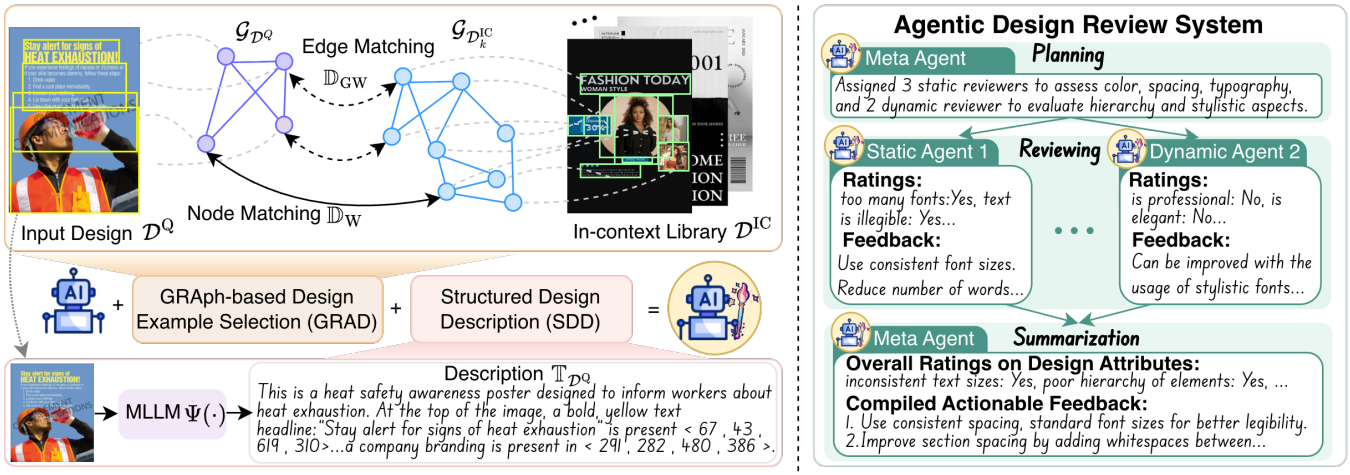

Figure 2: **Overview of the proposed design evaluation pipeline.** To systematically evaluate designs, GRAD constructs a graph representation of design elements using edge and node matching techniques, enabling structured retrieval of in-context design examples from a curated library (Sec 3). Structured Design Description (SDD) module generates design descriptions (Sec 3) to anchor the responses of each agent. The selected K designs, query design along with its description inform the review process, where a meta agent coordinates static and dynamic agents to assess disparate design attributes - part of the Agentic Design Review System. Static agents focus on a fixed set of attributes (e.g., typography), whereas dynamic agents evaluate the attributes which are contextualized to specific designs (e.g., stylistic qualities). The meta agent consolidates these insights into a final rating and provides actionable feedback for design improvements (Sec 3).

(i) Presence of bounding boxes: when layout metadata is available, bounding box coordinates $\mathcal{BB}^j$ are extracted for all elements in a design $\mathcal{D}^j$. Design components V^j are obtained by cropping the design using $\mathcal{BB}^j$: $V^j(i) = \text{Crop}(D^j, \mathcal{BB}^j(i)), ; \forall i \in 1, ..., N_{V^j}$, where N_{V^j} is the number of elements in $\mathcal{D}^j$ ($j \in Q, \text{IC}$), and $\text{Crop}(\cdot)$ denotes the cropping operation. Each cropped element is passed through CLIP to obtain embeddings: $\phi(V^j)(i) = \text{CLIP}(V^j(i)), \ \forall i \in 1, ..., N_{V^j}$. These embeddings of the corresponding design elements form the vertices $\mathbb{V} = \{\phi(V^j)(i)\}_{i=1}^{N_V}$ of the graph. The edges are composed of spatial and semantic distances, where spatial distance $d_{\text{spatial}}(u, v)$ between two elements (i.e., nodes of a graph) u and v is computed based on the normalized L_2 norm between the bounding box centroids $\mathcal{BB}^j(u)$ and $\mathcal{BB}^j(v)$ and the semantic distance $d_{\text{semantic}}(u, v)$ is computed as cosine distance between the respective embeddings $\mathbb{V}(u)$ and $\mathbb{V}(v)$. The combined edge weight is given as:

$$d^j(u, v) = d^j_{\text{spatial}}(u, v) + d^j_{\text{semantic}}(u, v) \tag{1}$$

(ii) Absence of bounding boxes: in the absence of layout metadata (e.g., for datasets w/o bounding boxes), we extract patch-level features from the final layers of a CLIP-ViT encoder. These local feature vectors form the graph: vertices $\mathbb{V}$ represent patch embeddings, and edges $\mathbb{E}$ encode semantic (cosine) distances between them. Lacking explicit structure, we rely solely on semantic relationships. Assuming an input design where inherently related elements are not grouped together, our approach GRAD will look for designs in the in-context library not only whose content is semantically close to the design under evaluation (via node matching) but also whose related elements are not clearly grouped (via edge matching). Therefore, GRAD enables structure-aware retrieval, improving the quality of selected in-context examples, achieved by the suitably combining Wasserstein (node matching) (Peyré, Cuturi et al. 2019; Cuturi 2013; Benamou et al. 2015; Luise et al. 2018) and Gromov-Wasserstein (edge matching) distances as described below. Given $\mathcal{D}^Q$, we select the most relevant in-context examples $\{\mathcal{D}^{\text{IC}}_k\}_{k=1}^K$ from $\mathcal{D}^{\text{IC}}$ using Wasserstein and Gromov-Wasserstein distances.

Wasserstein Distance (WD) calculates the pairwise distances between the node embedding sets $\mathbb{X}^Q$ and $\mathbb{Y}^k$ of the respective graphs $\mathcal{G}_{\mathcal{D}^Q}$ and $\mathcal{G}_{\mathcal{D}^{\text{IC}}_k}$. Considering two discrete distributions, $\psi^Q \in \mathbf{P}(\mathbb{X}^Q)$ and $\psi^k \in \mathbf{P}(\mathbb{Y}^k)$, where $\psi^Q = \sum_{i=1}^n z^Q(i)\delta_{x^Q}(i)$ and $\psi^k = \sum_{j=1}^m z^k(j)\delta_{y^k}(j)$; $\sum_i z^Q(i) = 1 = \sum_j z^k(j)$ where z^Q and z^k are the respective weight vectors for the probability distributions ψ^Q and ψ^k; and δ_{x^Q} is the Dirac-delta function centered on support point x^Q in the embedding space. The WD distance between ψ^Q and ψ^k is defined as:

$$\mathbb{D}_{\text{W}}(\mathcal{G}_{\mathcal{D}^Q}, \mathcal{G}_{\mathcal{D}^{\text{IC}}_k}) \triangleq \min_{\mathbf{\Phi} \in \mathbf{\Theta}(z^Q, z^k)} \sum_{i,j} \mathbf{\Phi}_{ij} \cdot c(x_i, y_j) \tag{2}$$

where $\mathbf{\Theta}(z^Q, z^k) = \{\mathbf{\Phi} \in \mathbb{R}_+^{n \times m} | \mathbf{\Phi}\mathbf{1}_m = z^Q, \mathbf{\Phi}^\top \mathbf{1}_n = z^k\}$, $c(x^Q(i), y^k(j))$ is cosine distance similarity metric, and $\mathbf{\Phi}$ is the transport plan, interpreting the amount of mass shifted from distributions $\psi^Q(i)$ to $\psi^k(j)$. An exact solution to the above expression leads to a sparse representation of transport plan $\mathbf{\Phi}$ with at most $(2 \cdot \max(m, n) - 1)$ non-zero

Algorithm 1: Overall Pipeline

Input: Query Design: $\mathcal{D}^Q$, In-context Designs: $\mathcal{D}^{\text{IC}}$, In-context samples to be retrieved: K, Static Attribute Buckets: B_{A_S}, Meta Agent: A_M, Static Agents: A_S, Number of Static Agents: N_S, World Attributes: $\mathcal{W}$, Bounding Boxes: $\mathcal{BB}$, CLIP Encoder: CLIP$(\cdot)$ Graph Construction Module: $\Omega(\cdot)$, Design Description module: $\Psi(\cdot)$, Wasserstein Distance func.: $\mathbb{D}_{\text{W}}(\cdot)$, Gromov-Wasserstein Distance function: $\mathbb{D}_{\text{GW}}(\cdot)$, Balancing factors: α, λ, Index Sorting function: **Argsort**$(\cdot)$, List Element Insertion function: **Insert**$(\cdot)$.

Output: Dynamic Attribute Buckets: B_{A_D}, Dynamic Agents: A_D, Number of Dynamic Agents: N_D, Feedback: $\mathcal{F}$, Attribute Rating: $\mathcal{R}$, Query Graph: $\mathcal{G}_{\mathcal{D}^Q}$, k^{th} in-context Graph: $\mathcal{G}_{\mathcal{D}^{\text{IC}}_k}$, Graph Dissimilarity Scores: $\mathbb{S}_L$, Design Description: $\mathbb{T}$.

$\mathcal{G}_{\mathcal{D}^Q} \leftarrow \Omega(\mathcal{D}^Q)$ ▷ graph construction, Sec 3
for $\mathcal{D}^{\text{IC}}_k \in \mathcal{D}^{\text{IC}}$ **do**
  $\mathcal{G}_{\mathcal{D}^{\text{IC}}_k} \leftarrow \Omega(\mathcal{D}^{\text{IC}}_k)$ ▷ graph construction, Sec 3
  $\mathbb{S}_{\text{l}} \leftarrow \alpha\mathbb{D}_{\text{W}}(\mathcal{G}_{\mathcal{D}^Q}, \mathcal{G}_{\mathcal{D}^{\text{IC}}_k}) + (1-\alpha)\mathbb{D}_{\text{GW}}(\mathcal{G}_{\mathcal{D}^Q}, \mathcal{G}_{\mathcal{D}^{\text{IC}}_k})$ ▷ Sec 3
  $\mathbb{S}_{\text{g}} \leftarrow 1 - \texttt{cos}(\text{CLIP}(\mathcal{D}^Q), \text{CLIP}(\mathcal{D}^{\text{IC}}_k))$ ▷ Sec 3
  $\mathbb{S}(\mathcal{D}^Q, \mathcal{D}^{\text{IC}}_k) \leftarrow \mathbb{S}_{\text{l}} + \mathbb{S}_{\text{g}}$ ▷ Sec 3
  $\mathbb{S}_L \leftarrow$ **Insert**$(\mathbb{S}_L, \mathbb{S}(\mathcal{D}^Q, \mathcal{D}^{\text{IC}}_k))$ ▷ inserting scores to list
$\mathbb{I}_S \leftarrow$ **Argsort**$(\mathbb{S}_L)$ ▷ index sorting
$\mathbb{I}_K \leftarrow \mathbb{I}_S[:K]$ ▷ Select top-*K* indices
$\mathbb{T}_{\mathcal{D}^Q} \leftarrow \Psi(\mathcal{D}^Q, \mathcal{BB}^Q)$ ▷ description generation, Sec 3
$A_D, B_{A_D}, N_D \leftarrow A_M(\mathcal{D}^Q, \mathbb{T}_{\mathcal{D}^Q}, A_S, B_{A_S}, N_S, \mathcal{W})$ ▷ planning
for $n_s \in 1, \cdots, N_S$ **do** ▷ reviewing, Sec 3
  $\mathcal{R}_S(n_s), \mathcal{F}_S(n_s) \leftarrow A_S^{(n_s)}(\mathcal{D}^Q, \mathbb{T}_{\mathcal{D}^Q}, \mathcal{D}^{\text{IC}}[\mathbb{I}_K], B_{A_S}^{(n_s)})$
for $n_s \in 1, ..., N_D$ **do** ▷ reviewing, Sec 3
  $\mathcal{R}_D(n_d), \mathcal{F}_D(n_d) \leftarrow A_D^{(n_d)}(\mathcal{D}^Q, \mathbb{T}_{\mathcal{D}^Q}, \mathcal{D}^{\text{IC}}[\mathbb{I}_K], B_{A_D}^{(n_d)})$
$\mathcal{R}, \mathcal{F} \leftarrow A_M([\mathcal{R}_S, \mathcal{R}_D], [\mathcal{F}_S, \mathcal{F}_D])$ ▷ summarization, Sec 3
return $\mathcal{R}, \mathcal{F}$

elements - this ensures an explainable and robust retrieval (De Goes et al. 2011).

Gromov-Wasserstein Distance (GWD) helps in matching the edges of the graphs and preserves graph topology by computing distances between pairs of nodes thereby ensuring inter-graph structural alignment (Peyré, Cuturi, and Solomon 2016; Alvarez-Melis and Jaakkola 2018). In the same discrete graph matching setting, GWD can be represented:

$$\mathbb{D}_{\text{GW}}(\mathcal{G}_{\mathcal{D}^Q}, \mathcal{G}_{\mathcal{D}^{\text{IC}}_k}) \triangleq \min_{\hat{\mathbf{\Phi}} \in \mathbf{\Theta}(z^Q, z^k)} \sum_{i,i',j,j'} \hat{\mathbf{\Phi}}_{ij} \hat{\mathbf{\Phi}}_{i'j'} \mathcal{L}\left(x_i^Q, y_j^k, x_{i'}^Q, y_{j'}^k\right) \tag{3}$$

where inter-graph structural similarity between two node pairs $(x_i^Q, x_{i'}^Q)$ and $(y_j^k, y_{j'}^k)$ is represented as $\mathcal{L}\left(x_i^Q, y_j^k, x_{i'}^Q, y_{j'}^k\right) = \|d^Q(x_i^Q, x_{i'}^Q) - d^k(y_j^k, y_{j'}^k)\|$, d^Q being the edge weight between a node pair in graph $\mathcal{G}_{\mathcal{D}^Q}$ (see Equation 1). Transport plan $\hat{\mathbf{\Phi}}$ is periodically updated to align the edges in different graphs belonging to disparate designs. A detailed set of steps is provided in Supp Algo 2.

The combined dissimilarity (since we are computing distances) score $\mathbb{S}_{\text{l}}$ is a weighted combination of Eqs. 2 and 3, based on which top-K designs are selected from the in-context design library: $\mathcal{D}^{\text{IC}} \subset \mathcal{D}$. $\mathbb{S}_{\text{l}} = \alpha\mathbb{D}_{\text{W}}(\mathcal{G}_{\mathcal{D}^Q}, \mathcal{G}_{\mathcal{D}^{\text{IC}}_k}) + (1-\alpha)\mathbb{D}_{\text{GW}}(\mathcal{G}_{\mathcal{D}^Q}, \mathcal{G}_{\mathcal{D}^{\text{IC}}_k})$. Along with $\mathbb{S}_{\text{l}}$ (on local representations), we add global scores on renditions: $\mathbb{S}_{\text{g}} = 1 - \texttt{cos}(\text{CLIP}(\mathcal{D}^Q), \text{CLIP}(\mathcal{D}^{\text{IC}}_k))$ where $\texttt{cos}(\cdot)$ is cosine similarity. The final expression becomes: $\mathbb{S}(\mathcal{D}^Q, \mathcal{D}^{\text{IC}}_k) = \mathbb{S}_{\text{l}} + \mathbb{S}_{\text{g}}$.

Structured Design Description (SDD)

For the input design $\mathcal{D}^Q$, our goal is to generate textual descriptions $\mathbb{T}_{\mathcal{D}^Q}$ containing description of elements (images, icons, texts, etc.) and how they are structured hierarchically as: $\mathbb{T}_{\mathcal{D}^Q} = \Psi\left(\mathcal{D}^Q, \mathcal{BB}^Q\right)$; if $\mathcal{BB}^Q \neq \emptyset$, else $\Psi\left(\mathcal{D}^Q\right)$. For example, a description may look like: "A title 'ABC' at the top $[bb_{11}^Q, bb_{12}^Q, bb_{13}^Q, bb_{14}^Q]$ with an image of 'X' below it $[bb_{21}^Q, bb_{22}^Q, bb_{23}^Q, bb_{24}^Q]$. Below the image, there is a text containing 'DEF', ... ". Passing both the graphic design and its textual description, optionally enriched with bounding box data $\mathcal{BB}^Q = bb_{ij}^Q$, enhances design attribute understanding and anomaly detection by combining visual input with structural, semantic, and relational context. These descriptions outperform raw metadata (e.g., xml/json; see Supp) by grounding MLLM responses in detailed visual context. This improves robustness across diverse layouts, facilitates clearer actionable feedback, and reduces hallucinations. We generate these descriptions by prompting an MLLM $\Psi(\cdot)$ (in Supp).

Agentic-Design Review System

Design literature (Graham 2002; Carpenter and Morin 2019; Williams 2007) emphasizes the importance of attributes like alignment, overlap, spacing. Any effective review system must rigorously evaluate these foundational principles to ensure design quality. To this end, we propose a structured agentic review framework, AGENTIC-DRS , with specialized *Static Agents* - predefined reviewers with fixed roles, focused on universally relevant design attributes. These agents serve as the backbone of the evaluation system, ensuring that essential design attributes are consistently reviewed. However, design evaluation is not solely a rule-based exercise; context-dependent attributes significantly influence perception. To capture these nuances, we introduce *Dynamic Agents* (evaluators/reviewers), which adapt based on the unique characteristics of a given design. These agents assess factors such as relative spacing, grouping, semantic effectiveness of communication, stylistic coherence, etc. which vary across different designs (i.e., dynamic in nature and contextualized on designs) and are not universally predefined.

The informed decision of which static and dynamic agents to activate for a particular query design is handled by a *Meta Agent*, which intelligently plans the evaluation process, ensuring that the most relevant aspects of a design are meticulously scrutinized - this constitutes the ***planning*** phase. Once individual agents have assessed the design and provided their ratings (***reviewing*** phase), the system must aggregate their insights into a coherent and actionable review. A consolidation mechanism, also managed by the Meta Agent, performs ***summarization***, synthesizing feedback, resolving potential inconsistencies, and generating a unified evaluation report. This ensures the final critique is holistic and actionable, guiding meaningful improvements in design, and not just a collection

Method	*Discrete Evaluation (Classification)*									*Continuous Evaluation (Correlation)*		
	Afixa			**Infographic**			**IDD**			**GDE (Haraguchi et al. 2024)**		
	Acc ↑	Sens ↑	Spec ↑	Acc ↑	Sens ↑	Spec ↑	Acc ↑	Sens ↑	Spec ↑	Alignment ↑	Overlap ↑	Whitespace ↑
Heuristic-based Evaluation	-	-	-	-	-	-	-	-	-	0.310	0.476	0.233
Gemini-1.5-Pro	59.45	62.11	60.18	54.88	55.61	54.85	64.37	67.89	65.59	0.586	0.759	0.641
Gemini-1.5-Pro + GRAD	62.19	64.08	62.45	56.21	60.72	56.18	68.21	68.65	67.16	0.623	0.778	0.676
Gemini-1.5-Pro + GRAD + SDD	65.62	67.95	66.31	59.76	63.84	59.05	69.09	68.94	70.42	0.671	0.783	0.691
AGENTIC-DRS $_{\text{Gemini-1.5-Pro}}$	72.17	74.94	70.85	65.97	67.41	68.58	75.43	**75.31**	76.22	0.712	0.821	0.739
$\Delta_{\text{AGENTIC-DRS - Gemini-1.5-Pro}}$	12.72 ↑	12.83 ↑	10.67 ↑	11.09 ↑	11.80 ↑	13.73 ↑	11.06 ↑	7.42 ↑	10.63 ↑	0.126 ↑	0.062 ↑	0.098 ↑
GPT-4o	62.91	65.42	64.26	58.26	61.92	56.74	65.72	65.38	66.57	0.597	0.782	0.665
GPT-4o + GRAD	64.57	68.65	65.18	60.41	63.57	59.66	68.51	67.26	70.85	0.639	0.796	0.688
GPT-4o + GRAD + SDD	67.33	69.60	68.21	64.95	66.21	62.12	70.16	69.44	73.92	0.677	0.809	0.703
AGENTIC-DRS $_{\text{GPT-4o}}$	**75.29**	**77.65**	**72.53**	**69.53**	**75.37**	**71.94**	**76.78**	74.56	**80.31**	**0.722**	**0.834**	**0.748**
$\Delta_{\text{AGENTIC-DRS - GPT-4o}}$	12.38 ↑	12.23 ↑	8.27 ↑	11.27 ↑	13.45 ↑	15.20 ↑	12.28 ↑	9.18 ↑	13.74 ↑	0.125 ↑	0.052 ↑	0.083 ↑

Table 1: **Performance of AGENTIC-DRS on the DRS-BENCH** . AGENTIC-DRS outperforms baseline methods by substantial margins in both evaluation protocols (*discrete* for attribute classification, and *continuous* for correlation with human labels).

of disjointed scores. Unlike prior works on design evaluation based on fixed heuristics (O'Donovan, Agarwala, and Hertzmann 2014) or end-to-end scoring (Goyal et al. 2025), our adaptive, peer-review-inspired framework accommodates evolving design principles and diverse styles.

We formally define the agents participating in the design evaluation process, comprising of Meta Agent A_M, Static Agents $A_{S_i} \forall i \in \{1, 2, ..., N_S\}$, and Dynamic Agents $A_{D_i} \forall i \in \{1, 2, ..., N_D\}$. $\mathcal{S}$ represents the state space describing the design under evaluation. The joint action space is denoted by $\mathbb{A}$, where each agent A_i executes actions a_i based on its policy π_i. The transition function $\mathcal{T} : \mathcal{S} \times \mathbb{A} \to \mathcal{S}$ models the evolution of the design evaluation state as agents contribute their assessments. For a given query design $\mathcal{D}^Q$, the agents collaborate in a structured sequence, via a directed interaction network (Fig 2) which can be decomposed into the aforementioned 3 phases (planning, reviewing, summarization), which we describe below.

(i) Planning: The meta-agent A_M plans the evaluation process by acting as a router which initiates the process by assigning static reviewers A_{S_i} based on predefined criteria and dynamic reviewers A_{D_i} based on attributes sampled from the open-world design principles (attributes) $\mathcal{W}$ which are deemed contextual and relevant for the design to be evaluated. Notably, each static evaluator A_{S_i} is responsible for evaluating a specific set of design attributes with predefined buckets (in Supp) as $B_{A_S}(i) \triangleq \{w^{(i)}_{A_S}(1), ..., w^{(i)}_{A_S}(k)\},\ w^{(i)}_{A_S}(k) \in W_{A_S} \subset \mathcal{W},\ i \in \{1, ..., N_S\}$, which are embedded in their respective prompts $p^{(i)}_{A_S}(B_{A_S}(i)) \in \mathcal{P}$ (the collective prompt space is represented as $\mathcal{P}$). Whereas, for each dynamic evaluator, attributes are first sampled $W_{A_D} \subseteq \mathcal{W} - W_{A_S}$ and dynamically bucketed on the fly (as decided by the meta agent) into N_D buckets where each $B_{A_D}(i) \triangleq \{w^{(i)}_{A_D}(1), ..., w^{(i)}_{A_S}(l)\},\ w^{(i)}_{A_D}(l) \in W_{A_D},\ i \in \{1, ..., N_D\}$ followed by dynamic prompt creation $p^{(i)}_{A_D} \in \mathcal{P}$.

$$a_M \sim \pi_M(\cdot|s, \mathcal{D}^Q, \mathcal{D}^{\text{IC}}_K), \quad s \sim \mathcal{S} \tag{4}$$

(ii) Reviewing: Both groups of agents (static and dynamic) follow a policy π_i, determining its output based on the previous agent (i.e., meta agent), and the top-K n-context (IC) samples $\mathcal{D}^{\text{IC}}_K$ from the in-context Data Library $\mathcal{D}^{\text{IC}}$ ($\mathcal{D}^{\text{IC}}_K \subseteq \mathcal{D}^{\text{IC}} \subset \mathcal{D}$), in the interaction network:

$$a_i \sim \pi_i(\cdot|s, \{a_M\}_{A_M}, \mathcal{D}^Q, \mathcal{D}^{\text{IC}}_K), \quad s \sim \mathcal{S} \tag{5}$$

where A_M's outputs influence A_i. The aggregated actions $a = (a_1, \ldots, a_N)$ collectively determine the next state: $s_{t+1} = \mathcal{T}(s_t, a_t) = \text{Concat}(s_t, a_t)$, where Concat represents the concatenation operation that updates the design evaluation context. These agents assign *quantitative* measures (design attribute ratings) and *qualitative* measures (actionable feedback), respectively denoted by $\mathcal{R}_S$ and $\mathcal{F}_S$ (for static agents), and $\mathcal{R}_D$ and $\mathcal{F}_D$ (for dynamic agents).

(iii) Summarization: The meta-agent A_M collates the scores and integrates the feedback it receives from all static and dynamic reviewers to construct a final evaluation.

$$a_M \sim \pi_M(\cdot|s, \{a_j\}_{A_j \in \text{Pred}(A_M)}), \quad s \sim \mathcal{S} \tag{6}$$

where $\text{Pred}(A(i))$ represents predecessor agent(s) who directly influence the outputs of the successor agent(s). Using this mechanism, redundant feedback is removed, refining the assessment and obtaining a final list of actionable feedback $\mathcal{F}$ (from $\mathcal{F}_S$, $\mathcal{F}_D$) and Attribute Ratings $\mathcal{R}$ (from $\mathcal{R}_S$, $\mathcal{R}_D$).

Overall Framework

We summarize the overall flow of our pipeline in Algo 1. Our key novelties are: (*i*) to adapt the graph matching algorithm based on localized representations (Supp Algos 1 - 2) for selecting the top-K designs based on scores (lines 1 - 9, Algorithm 1), (*ii*) to anchor MLLM responses via structured design descriptions generated using design renditions and bounding boxes (line 10, Algo 1), and (*iii*) to introduce agentic design review framework comprising of meta, static and dynamic agents and involving planning, reviewing and summarization mechanisms (lines 11 - 16, Algo 1).

4 Experiments and Results

DRS-BENCH : Design Evaluation Benchmark

Graphic design principles support clear compositions (Carpenter and Morin 2019; Graham 2002; Williams 2007). Existing heuristic methods (O'Donovan, Agarwala, and Hertzmann 2014; Goyal et al. 2025) evaluate limited aspects and overlook color, typography, and aesthetics. Despite advances

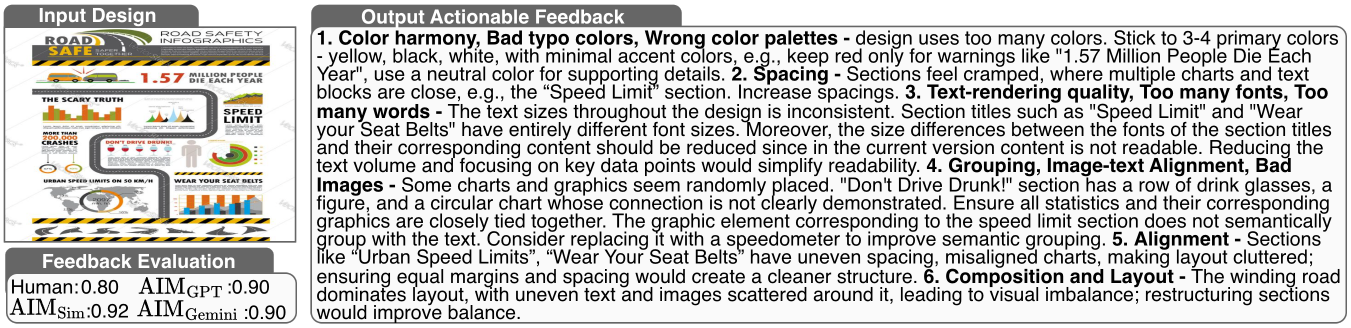

Figure 3: **Generated feedback along with the design attributes** which are found to be inconsistent for the input design evaluated (best viewed when zoomed). Feedback evaluation scores are also reported (as outlined in Sec 4).

Method	Metric	GDE	Infographic	Afixa	IDD
AGENTIC-DRS $_{Gemini}$	Human	0.744	0.682	0.720	0.742
AGENTIC-DRS $_{Gemini}$	AIM$_{Sim}$	0.851	0.806	0.821	0.834
AGENTIC-DRS $_{Gemini}$	AIM$_{Gemini}$	0.792	0.758	0.769	0.802
AGENTIC-DRS $_{Gemini}$	AIM$_{GPT\text{-}4o}$	0.795	0.736	0.762	0.785
AGENTIC-DRS $_{GPT\text{-}4o}$	Human	0.762	0.708	0.736	0.740
AGENTIC-DRS $_{GPT\text{-}4o}$	AIM$_{Sim}$	0.881	0.835	0.817	0.837
AGENTIC-DRS $_{GPT\text{-}4o}$	AIM$_{Gemini}$	0.829	0.774	0.794	0.791
AGENTIC-DRS $_{GPT\text{-}4o}$	AIM$_{GPT\text{-}4o}$	0.832	0.763	0.803	0.804

Table 2: **Feedback evaluation of AGENTIC-DRS** (both GPT-4o and Gemini). Strong correspondences observed across different AIM variations and with human ratings.

in automated design generation (Luo et al. 2024; Inoue et al. 2023a; Yamaguchi 2021; Yang et al. 2024; Tang et al. 2023; Lin et al. 2023; Guerreiro et al. 2024), no standard evaluation exists. To address such gaps, we propose DRS-BENCH , a unified benchmark for evaluating design effectiveness.

Attributes. A core aspect of our benchmark is the multi-dimensional evaluation of designs through a comprehensive set of design attributes, collectively termed "World Attributes" ($\mathcal{W}$). These attributes capture fundamental design flaws and best practices. *Text-rendering quality* assesses legibility based on font size, weight, and clarity. *Too many words* penalizes cluttered text, while *too many fonts* and *bad typography colors* flag inconsistent style, poor contrast. *Composition and layout*, *alignment*, and *spacing* ensure visual balance, favoring structured designs with clear margins and uniform spacing. *Color harmony* and *wrong color palettes* measure the effectiveness of color choices in maintaining aesthetic appeal. *Style* evaluates stylistic aspects and consistency, while *grouping* and *image-text alignment* address content organization, readability. *Aesthetics* considers overall visual appeal, penalizing *overlap*, *bad images*, poorly arranged elements.

Datasets. In DRS-BENCH , we include 4 datasets which are: *GDE:* DRS-BENCH leverages publicly available GDE dataset (Haraguchi et al. 2024), hosting a large collection of 700 banner and poster designs each of which containing 3 attributes: alignment, overlap, and white space, on a 1 to 10 scale. Layout metadata information are not publicly available. *Afixa:* It consists of 71 designs, in DRS-BENCH as collected from a public platform Roboflow which consists of *yes* or *no* values for 5 design attributes: *wrong color palettes pairings*, *bad typo colors*, *bad images*, *too many words*, *too many fonts*. Layout metadata information are not available.

Infographic: It consists of 55 samples collected from the Roboflow platform. This consists of the layout metadata information (xml) for the elements (i.e., bounding boxes for the elements). Each design has *yes* or *no* responses for all the 15 attributes (refer to Supp and Sec 4).

Internal Design Dataset (IDD): We internally collect 137 design samples each of which has *yes* or *no* responses to the 15 design attributes (see Supp and Sec 4). The designs are professionally curated and typically comprises of flyers, invitations, posters, albeit with some inconsistencies in design attributes. IDD consists of the layout metadata information (xml) with bounding boxes for the design elements.

Evaluation Metrics

Attribute Evaluation: Design flaws often coexist, impacting readability, style, and aesthetics. Using multiple labels enables more detailed evaluation than a single *goodness-of-fit* score. For Afixa, Infographic, and IDD , attributes are labeled *yes*/*no*, forming a multi-label classification task. We report mean Sensitivity, Specificity, and Accuracy as *Discrete Evaluation*. For GDE, where attributes are rated on a 1–10 scale, we report Pearson correlation with human ratings (Haraguchi et al. 2024), termed *Continuous Evaluation*.

Feedback Evaluation: To evaluate feedback quality, we use the Actionable Insights Metric (AIM): (*i*) AIMGPT-4o/AIMGem, where GPT-4o/Gemini rates how well feedback addresses ground truth problems (converted to sentences), (*ii*) AIM$_{Sim}$ measures semantic similarity between problems and feedback using the GTE-L (Li et al. 2023).

Baselines

To the best of our knowledge, there does not exist any design-aware MLLMs let alone agentic frameworks. To compare AGENTIC-DRS , we introduce baselines using two strong MLLMs—Gemini-1.5 Pro (Team et al. 2024) and GPT-4o (Hurst et al. 2024), augmented with design-aware components like GRAD and SDD. These baselines use only a single MLLM agent. We also include the vanilla GPT-4o implementation from Haraguchi et al. (2024) (used in Table 1) and report heuristic results (O'Donovan, Agarwala, and Hertzmann 2014) on the GDE dataset (Haraguchi et al. 2024). However,

In-context Sampling Method	GDE (Avg.)	Infographic (Acc)	Afixa (Acc)
Random Selection (9)	0.743	63.75	72.06
Global Features (6)	0.752	64.39	72.94
Description-based (6)	0.755	67.05	73.98
GRAD (w/o Global Features) (5)	0.760	67.13	74.81
GRAD (4)	**0.768**	**69.53**	**75.29**

Table 3: **Ablation on different in-context design selection methods** for AGENTIC-DRS $_{\text{GPT-4o}}$. Best results are obtained with GRAD, preserving semantic, spatial and structural information. Best K for top-K retrieval are provided in brackets beside method.

GRAD	SDD		Infographic (Acc)	IDD (Acc)
	w/ Bnd. Box	w/o Bnd. Box		
✗	✗	✗	63.11	69.75
✗	✓	✗	64.52	71.03
✗	✗	✓	65.71	72.86
✓	✗	✗	66.48	73.95
✓	✓	✗	67.82	74.57
✓	✓	✓	**69.53**	**76.78**

Table 4: **Effect of different components** of AGENTIC-DRS $_{\text{GPT-4o}}$. Results on dataset with layout metadata information are reported. Substantial improvements are found upon adding GRAD and SDD.

α	Infog. (Acc)	Afixa (Acc)
0	68.10	73.95
1	66.85	72.86
0.25	68.94	74.38
0.75	67.21	73.14
0.5	**69.53**	**75.29**

Table 5: **Ablation on** α for AGENTIC-DRS $_{\text{GPT-4o}}$ on Infographic and Afixa datasets.

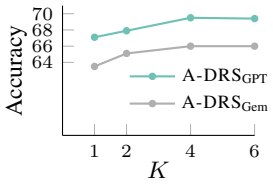

Figure 4: **Impact of** K **in GRAD** for AGENTIC-DRS .

Afixa, Infographic, and IDD include design attributes beyond conventional ones (e.g., alignment, overlap, spacing) and are thus not compatible with heuristic-based evaluation.

Results

The Table 1 presents quantitative comparisons of AGENTIC-DRS against baseline approaches on DRS-BENCH . Each MLLM-based evaluation reports mean of five independent runs. For in-context library, we use 5, 15, 20, 25% data respectively from GDE, IDD, Afixa, Infographic datasets during evaluations. Results highlight that incorporating design awareness and multi-agent interaction significantly outperforms single-agent systems in both Discrete (multi-attribute classification) and Continuous (correlation with human labels) evaluations. GPT-4o consistently outperforms Gemini-1.5-Pro across datasets and metrics (except Sensitivity on IDD), aligning with prior findings (Qi et al. 2023). On GDE, all MLLM-based evaluations achieve significant gains over heuristic-based methods, corroborating Haraguchi et al. (2024). Table 2 reports qualitative scores for AGENTIC-DRS Gemini and AGENTIC-DRS GPT-4o on four datasets, using the three feedback evaluation metrics from Sec 4. We include Human scores (averaged over five raters, normalized to [0, 1]) for reference. AIM$_{\text{Sim}}$, AIM$_{\text{Gemini}}$, and AIM$_{\text{GPT-4o}}$ scores strongly correlate with Human ratings, validating effectiveness of AIM for actionable feedback evaluation. Fig. 3 shows a qualitative example from the Infographics dataset, illustrating how AGENTIC-DRS delivers actionable design feedback. Feedback quality is supported by AIM scores (Sec 4) and (normalized) human ratings. GRAD selects relevant examples for better design understanding, while SDD grounds element positions, enabling precise references (e.g., inconsistent spacing in the "Speed Limit" section), improving feedback *actionability*. The stylistic aspects (elegance, minimality, etc.) are effectively handled by the dynamic experts, as spawned by the meta agent.

5 Discussions and Analysis

Analyzing the Impact of GRAD. Table 3 compares GRAD with four retrieval methods. *Random* exemplar selection performs worst on DRS-BENCH . CLIP *global features* improve results over random, aligning with prior work (Ferber et al. 2024). A *description-based* method using SDD-generated text (Sec 3) outperforms global features, especially when element coordinates are available (e.g., Infographic). Removing global features from GRAD causes a slight drop, confirming the value of combining structural and global cues. GRAD achieves the best performance with fewer in-context samples by preserving spatial and compositional structure rather than treating designs as unordered features. Figure 4 shows similar trends in Accuracy across K values for GPT-4o and Gemini on Infographic, confirming GRAD's robustness. Table 4 shows GRAD significantly boosts performance even without SDD (see Table 1 for single-agent baselines).

Assessing the Role of SDD. Design renditions capture aesthetic and spatial patterns, while textual descriptions provide an anchor for understanding. Table 4 shows that even without GRAD, SDD significantly boosts evaluation performance in AGENTIC-DRS , with similar trends in Table 1 for single MLLM models (GPT/Gemini). Performance improves further with bounding boxes, demonstrating that detailed visual descriptions with coordinate references (akin to phrase grounding or referring expressions) enhance model comprehension and anomaly detection in designs. We use structured descriptions over raw XML metadata because LLMs typically process textual information more effectively (see Supp).

Analyzing the Effect of α. Table 5 shows the impact of different α values in balancing WD and GWD. Using only GWD ($\alpha = 0$) outperforms WD alone ($\alpha = 1$), but the best performance occurs at $\alpha = 0.5$, denoting the need for coupling semantic feature matching with structural similarities.

6 Conclusion

We present the first Agentic framework for multi-dimensional critique of graphic designs, producing both scores and actionable feedback. This can aid novice designers and support evaluation of generative model design outputs (Jia et al. 2023b; Inoue et al. 2024). Our novel exemplar selection and prompt expansion methods are central to the framework's effectiveness, as shown through extensive experiments. A key next step is automating actionable feedback application, enabling a self-improving graphic design generation system.

References

Alvarez-Melis, D.; and Jaakkola, T. S. 2018. Gromov-Wasserstein alignment of word embedding spaces. *arXiv preprint arXiv:1809.00013*.

Benamou, J.-D.; Carlier, G.; Cuturi, M.; Nenna, L.; and Peyré, G. 2015. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2): A1111–A1138.

Carpenter, R.; and Morin, C. 2019. Thinking Like a Designer. *Multimodal Composing: Strategies for Twenty-First-Century Writing Consultations*, 67.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

De Goes, F.; Cohen-Steiner, D.; Alliez, P.; and Desbrun, M. 2011. An optimal transport approach to robust reconstruction and simplification of 2d shapes. In *Computer Graphics Forum*, volume 30, 1593–1602. Wiley Online Library.

Dou, Q.; Zheng, X. S.; Sun, T.; and Heng, P. 2019. Webthetics: Quantifying webpage aesthetics with deep learning. *Int. J. Hum. Comput. Stud.*, 124: 56–66.

Ferber, D.; Wölflein, G.; Wiest, I. C.; Ligero, M.; Sainath, S.; Ghaffari Laleh, N.; El Nahhas, O. S.; Müller-Franzes, G.; Jäger, D.; Truhn, D.; et al. 2024. In-context learning enables multimodal large language models to classify cancer pathology images. *Nature Communications*, 15(1): 10104.

Goyal, S.; Mahajan, A.; Mishra, S.; Udhayanan, P.; Shukla, T.; Joseph, K.; and Srinivasan, B. V. 2025. Design-o-meter: Towards Evaluating and Refining Graphic Designs.

Graham, L. 2002. Basics of Design: Layout and Typography for Beginners.(2ndedn). *New York*.

Guerreiro, J. J. A.; Inoue, N.; Masui, K.; Otani, M.; and Nakayama, H. 2024. LayoutFlow: flow matching for layout generation. In *European Conference on Computer Vision*, 56–72. Springer.

Haraguchi, D.; Inoue, N.; Shimoda, W.; Mitani, H.; Uchida, S.; and Yamaguchi, K. 2024. Can GPTs Evaluate Graphic Design Based on Design Principles? In Igarashi, T.; and Hu, R., eds., *SIGGRAPH Asia 2024 Technical Communications, SA 2024, Tokyo, Japan, December 3-6, 2024*, 5:1–5:4. ACM.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Inoue, N.; Kikuchi, K.; Simo-Serra, E.; Otani, M.; and Yamaguchi, K. 2023a. Layoutdm: Discrete diffusion model for controllable layout generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10167–10176.

Inoue, N.; Kikuchi, K.; Simo-Serra, E.; Otani, M.; and Yamaguchi, K. 2023b. Towards flexible multi-modal document models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14287–14296.

Inoue, N.; Masui, K.; Shimoda, W.; and Yamaguchi, K. 2024. OpenCOLE: Towards Reproducible Automatic Graphic Design Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8131–8135.

Jia, P.; Li, C.; Liu, Z.; Shen, Y.; Chen, X.; Yuan, Y.; Zheng, Y.; Chen, D.; Li, J.; Xie, X.; Zhang, S.; and Guo, B. 2023a. COLE: A Hierarchical Generation Framework for Graphic Design. *CoRR*, abs/2311.16974.

Jia, P.; Li, C.; Yuan, Y.; Liu, Z.; Shen, Y.; Chen, B.; Chen, X.; Zheng, Y.; Chen, D.; Li, J.; et al. 2023b. COLE: A Hierarchical Generation Framework for Multi-Layered and Editable Graphic Design. *arXiv preprint arXiv:2311.16974*.

Kong, W.; Jiang, Z.; Sun, S.; Guo, Z.; Cui, W.; Liu, T.; Lou, J.; and Zhang, D. 2023. Aesthetics++: Refining Graphic Designs by Exploring Design Principles and Human Preference. *IEEE Trans. Vis. Comput. Graph.*, 29(6): 3093–3104.

Laurençon, H.; Tronchon, L.; and Sanh, V. 2024. Unlocking the conversion of Web Screenshots into HTML Code with the WebSight Dataset. *CoRR*, abs/2403.09029.

Levi, E.; Brosh, E.; Mykhailych, M.; and Perez, M. 2023. Dlt: Conditioned layout generation with joint discrete-continuous diffusion layout transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2106–2115.

Li, Z.; Zhang, X.; Zhang, Y.; Long, D.; Xie, P.; and Zhang, M. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Lin, J.; Guo, J.; Sun, S.; Yang, Z.; Lou, J.-G.; and Zhang, D. 2023. Layoutprompter: awaken the design ability of large language models. *Advances in Neural Information Processing Systems*, 36: 43852–43879.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Luise, G.; Rudi, A.; Pontil, M.; and Ciliberto, C. 2018. Differential properties of sinkhorn approximation for learning with wasserstein distance. *Advances in Neural Information Processing Systems*, 31.

Luo, C.; Shen, Y.; Zhu, Z.; Zheng, Q.; Yu, Z.; and Yao, C. 2024. Layoutllm: Layout instruction tuning with large language models for document understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15630–15640.

Ngo, D. C. L.; Teo, L. S.; and Byrne, J. G. 2000. A mathematical theory of interface aesthetics. In *Visual mathematics*, volume 2. Mathematical Institute SASA.

O'Donovan, P.; Agarwala, A.; and Hertzmann, A. 2014. Learning layouts for single-pagegraphic designs. *IEEE transactions on visualization and computer graphics*, 20(8): 1200–1213.

Peyré, G.; Cuturi, M.; and Solomon, J. 2016. Gromov-wasserstein averaging of kernel and distance matrices. In *International conference on machine learning*, 2664–2672. PMLR.

Peyré, G.; Cuturi, M.; et al. 2019. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6): 355–607.

Qi, Z.; Fang, Y.; Zhang, M.; Sun, Z.; Wu, T.; Liu, Z.; Lin, D.; Wang, J.; and Zhao, H. 2023. Gemini vs gpt-4v: A preliminary comparison and combination of vision-language models through qualitative cases. *arXiv preprint arXiv:2312.15011*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Sawada, S.; Suzuki, T.; Yamaguchi, K.; and Toyoda, M. 2024. Visual Explanation for Advertising Creative Workflow. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–8.

Shridhar, M.; Thomason, J.; Gordon, D.; Bisk, Y.; Han, W.; Mottaghi, R.; Zettlemoyer, L.; and Fox, D. 2020. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 10737–10746. Computer Vision Foundation / IEEE.

Si, C.; Zhang, Y.; Yang, Z.; Liu, R.; and Yang, D. 2024. Design2Code: How Far Are We From Automating Front-End Engineering? *CoRR*, abs/2403.03163.

Son, K.; Choi, D.; Kim, T. S.; and Kim, J. 2024. Demystifying tacit knowledge in graphic design: Characteristics, instances, approaches, and guidelines. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–18.

Tabata, S.; Yoshihara, H.; Maeda, H.; and Yokoyama, K. 2019. 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*, 9:1–9:2. ACM.

Tang, Z.; Wu, C.; Li, J.; and Duan, N. 2023. Layoutnuwa: Revealing the hidden layout expertise of large language models. *arXiv preprint arXiv:2309.09506*.

Team, G.; Georgiev, P.; Lei, V. I.; Burnell, R.; Bai, L.; Gulati, A.; Tanzer, G.; Vincent, D.; Pan, Z.; Wang, S.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Williams, R. 2007. The Non-Designer's Design and Type Books, Deluxe Edition.

Xie, J.; Chen, Z.; Zhang, R.; Wan, X.; and Li, G. 2024. Large Multimodal Agents: A Survey. *CoRR*, abs/2402.15116.

Yamaguchi, K. 2021. Canvasvae: Learning to generate vector graphic documents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5481–5489.

Yang, T.; Luo, Y.; Qi, Z.; Wu, Y.; Shan, Y.; and Chen, C. W. 2024. Posterllava: Constructing a unified multi-modal layout generator with llm. *arXiv preprint arXiv:2406.02884*.

Zen, M.; and Vanderdonckt, J. 2014. Towards an evaluation of graphical user interfaces aesthetics based on metrics. In Bajec, M.; Collard, M.; and Deneckère, R., eds., *IEEE 8th International Conference on Research Challenges in Information Science, RCIS 2014, Marrakech, Morocco, May 28-30, 2014*, 1–12. IEEE.

Zhang, P.; Dong, X.; Zang, Y.; Cao, Y.; Qian, R.; Chen, L.; Guo, Q.; Duan, H.; Wang, B.; Ouyang, L.; et al. 2024. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*.

Zhang, Y.; Maezawa, A.; Xia, G.; Yamamoto, K.; and Dixon, S. 2023. Loop Copilot: Conducting AI Ensembles for Music Generation and Iterative Editing. *CoRR*, abs/2310.12404.

Zhao, N.; Cao, Y.; and Lau, R. W. 2018a. What characterizes personalities of graphic designs? *ACM Transactions on Graphics (TOG)*, 37(4): 1–15.

Zhao, N.; Cao, Y.; and Lau, R. W. H. 2018b. What characterizes personalities of graphic designs? *ACM Trans. Graph.*, 37(4): 116.