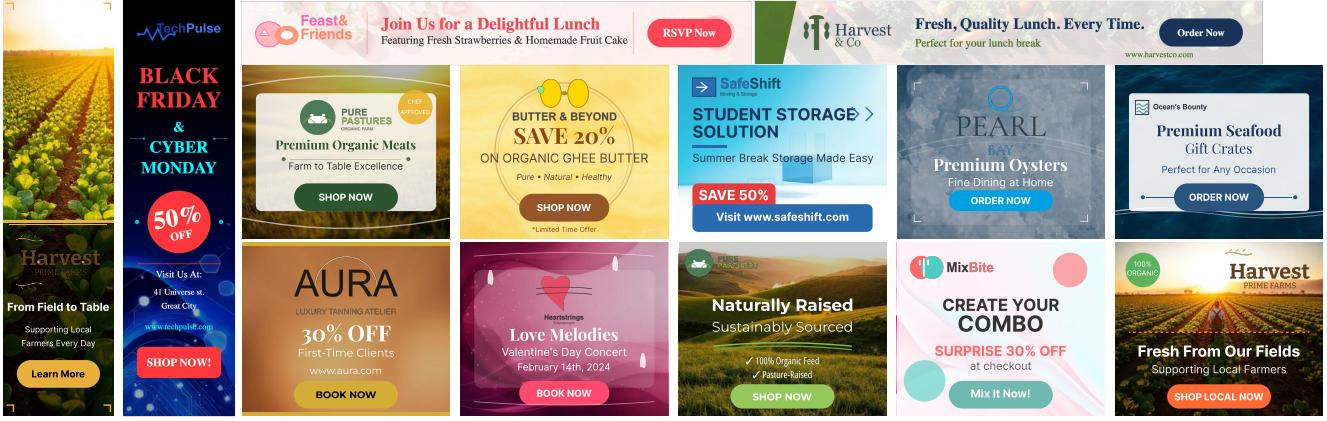


# BannerAgency: Advertising Banner Design with Multimodal LLM Agents

Heng Wang<sup>†</sup>, Yotaro Shimose, Shingo Takamatsu  
Sony Group Corporation



(a) BannerAgency with SVG implementation.

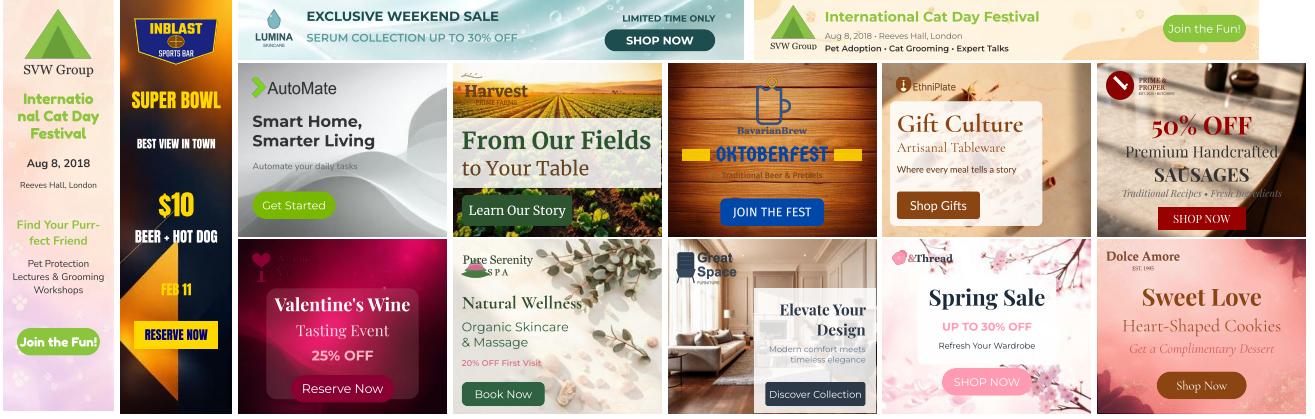


Figure 1. BannerAgency supports designs of various sizes including three common advertisement sizes: skyscraper ( $160 \times 600$ ), leaderboard ( $728 \times 90$ ), medium rectangle ( $300 \times 250$ ). It supports direct SVG generation or rendering in Figma. Both enable continuing editing at element level. Figures best viewed in Adobe Acrobat.

## Abstract

*Advertising banners are critical for capturing user attention and enhancing advertising campaign effectiveness. Creating aesthetically pleasing banner designs while conveying the campaign messages is challenging due to the large search space involving multiple design elements. Addition-*

*ally, advertisers need multiple sizes for different displays and various versions to target different sectors of audiences. Since design is intrinsically an iterative and subjective process, flexible editability is also in high demand for practical usage. While current models have served as assistants to human designers in various design tasks, they typically handle only segments of the creative design process or produce pixel-based outputs that limit editability. This paper introduces a training-free framework for fully automated*

<sup>†</sup>Corresponding author: Heng Wang (heng.wang@sony.com)

*banner ad design creation, enabling frontier multimodal large language models (MLLMs) to streamline the production of effective banners with minimal manual effort across diverse marketing contexts. We present BannerAgency, an MLLM agent system that collaborates with advertisers to understand their brand identity and banner objectives, generates matching background images, creates blueprints for foreground design elements, and renders the final creatives as editable components in Figma or SVG formats rather than static pixels. To facilitate evaluation and future research, we introduce BannerRequest400, a benchmark featuring 100 unique logos paired with 400 diverse banner requests. Through quantitative and qualitative evaluations, we demonstrate the framework’s effectiveness, emphasizing the quality of the generated banner designs, their adaptability to various banner requests, and their strong editability enabled by this component-based approach.*

## 1. Introduction

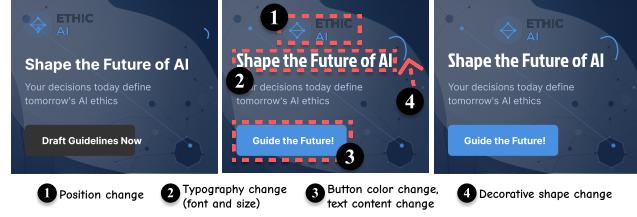
Advertising banners constitute an instrumental medium in digital marketing campaigns. These graphical displays efficiently capture viewer attention, communicate strategic messaging, and enhance brand visibility across competitive market environments [10, 37, 55, 60]. Unlike generic images, banner ad images are visually harmonized compositions of multiple creative assets including backdrops, brand-related logos or product images, click-to-action (CTA) buttons, campaign texts, and other decorative elements such as shapes and patterns [49, 61]. Campaign texts further require typography design choices, contributing to a vast search space that demands significant time and effort [16, 63]. In addition, advertisers have various size requirements for different display needs [5] and diverse design requirements tailored to specific target audiences and purposes [15]. As design remains inherently creative and subjective, ensuring human interaction and editing capability is crucial for practical applications [21, 33].

Current models accelerate this banner design process by automating separate aspects: product background generation [12, 16, 19, 63, 68], template-based layout matching [3, 8, 18, 39, 69], layout generation [25, 49, 61, 73], and automatic typography design [14, 28, 54]. Advanced scene-text rendering models [11, 45, 59, 74] and text-to-image (T2I) models such as DALL-E3 [6] are suboptimal for banner design due to their pixel-based generation paradigm, which produces outputs that are difficult to edit and often contains incoherent text as in Fig. 2a. Although TypeR [58] addresses the “typos” from T2I models through detection and correction of incorrectly rendered text, it still constrains the design process to a pixel-based approach, limiting subsequent editability. Recent works [31, 33, 64] explore automatic poster generation from prompts, but banner de-

Method	DALL-E3	OpenCOLE	BannerAgency
Sample			
Multimodal Input	✗	✗	✓
Size Variation	✗	✗	✓
Editability	✗	○	✓

✗: not supported    ○: partially supported.    ✓: supported.

(a) Left: multimodal banner request. Right: Algorithm comparison.

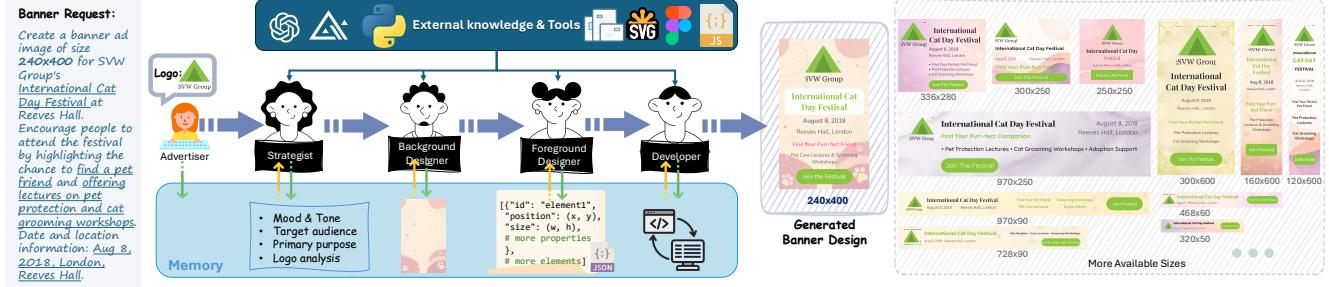


(b) An example of element-level editing.

Figure 2. Algorithm comparison and BannerAgency’s editability.

sign requires handling both structured requests and unstructured brand identity like logos. Their data dependency and fine-tuning further limit adaptability to diverse banner sizes and requests. Although these works allow text modification through typography rendering, they merge decorative elements with images, restricting editability (see Fig. 2a, where COLE-based[31, 33] output’s black rectangle cannot be removed).

To bridge these gaps and deliver a fully automatic yet editable banner design solution, we introduce BannerAgency, a training-free framework powered by multimodal large language models (MLLMs) [4, 44, 52]. Our fundamental idea is to utilize MLLMs as AI agents [17, 67] that simulate how a human design team works collaboratively from conception to implementation [15], leveraging the emergent design knowledge inherited from web-scale training data [27, 41, 65]. With tool calling and memory sharing capabilities of LLM agents [22, 26, 47], we effectively navigate the extensive design search space through our specialized agent architecture. The system comprises: (1) a *Strategist* that constrains design possibilities by establishing brand guidelines and campaign objectives; (2) a *Background Designer* responsible for visual backdrop generation; (3) a *Foreground Designer* that creates blueprints for banner assets with engaging content and harmonized styling and positioning; and (4) a *Developer* that transforms these specifications into editable components via Figma plugin code or SVG generation (see Fig. 1 for visual examples of the final designs). Unlike pixel-based methods that produce raster outputs, BannerAgency generates designs directly within professional tools, enabling seamless editing of major design components—text, typography, but-



**Figure 3. BannerAgency overview.** Given a logo and request, BannerAgency begins with the *Strategist* analyzing banner objectives, followed by the *Background Designer* creating matching backgrounds, then the *Foreground Designer* producing element blueprints, and concludes with the *Developer* rendering the final design as editable components. With access to external knowledge, tool-calling capabilities, and shared memory, BannerAgency enables context-aware, harmonized decisions and supports multiple banner sizes if requested.

tons, shapes, and layouts—within a familiar interface used by professional designers as illustrated in Fig. 2b.

To catalyze further exploration to advertising banner creation, we introduce BannerRequest400, a dedicated benchmark featuring 100 uniquely created logos paired with 400 diverse banner requests spanning multiple target audiences and primary purposes. Both quantitative and qualitative assessments confirm BannerAgency’s capabilities in generating high-quality banner designs that effectively adapt to various requirements while offering unparalleled editing flexibility through its component-based output format.

Our contributions can be summarized as follows:

- We propose BannerAgency, a training-free MLLM agent system that processes minimal multimodal inputs (brand logos, advertiser requirements) and autonomously generates all necessary creative assets as harmonized banner designs.
- Simulating professional design workflows—from strategy development to technical implementation, BannerAgency leverages MLLMs’ extensive world knowledge to navigate the complex design space while ensuring complete editability through component-based implementation in design tools such as Figma or as SVG format.
- To catalyze research in agent-driven advertising banner design, we introduce BannerRequest400, a curated benchmark featuring 100 uniquely created logos paired with 400 diverse banner requests spanning multiple target audiences and primary purposes.
- Through comprehensive evaluation, we exhibit BannerAgency’s effectiveness in generating high-quality, versatile, and editable banner designs that can serve either as final outputs or as starting points for subjective refinement.

## 2. Related Work

### 2.1. Graphic Design Generation

Current approaches to graphic design generation face significant limitations. Pixel-based text-to-image models [6, 7,

56, 72] require cumbersome editing, while layout generation research [13, 29, 30, 32, 38, 42, 43, 53, 57, 66, 76] has not adequately addressed simultaneous element generation and arrangement [35]. Existing solutions in prompt-based poster generation [31, 33, 64] rely on domain-specific training, accept only unimodal text input (not multimodal inputs like logos), and lack flexibility. While LayoutPrompt [41] offers a training-free approach, it addresses only layout without end-to-end capabilities. We propose a training-free agent framework that leverages MLLMs to simultaneously generate and arrange design elements in editable code representation from multimodal inputs, enabling flexible banner creation without domain-specific fine-tuning.

### 2.2. Multimodal LLM Agents

MLLMs [1, 40, 44, 77] have recently spurred an emerging wave of multimodal agent research [17, 24, 67]. These agents—autonomous entities that leverage tool-calling capabilities, planning, memory, and goal-directed reasoning—have been applied to various design domains including presentation slides generation [20, 75], image manipulation [65, 70], infographic design [9, 27], and sketch creation [62]. Our work extends this evolving frontier by leveraging MLLMs’ emergent capabilities to pioneer a comprehensive banner design solution that generates high-quality designs adaptable to diverse banner requests while ensuring strong editability through its component-based approach.

## 3. BannerAgency

Taking a festival promotion banner request for example, Fig. 3 illustrates the overall workflow of BannerAgency. Simulating how graphic designers work in banner ad industry, our BannerAgency consists of a *Strategist* to communicate with advertisers on their requirements, a *Background Designer* to draw the visual canvas, a *Foreground Designer* to craft the foreground elements such as logo, text, CTA button, and decorative elements, and a *Developer* to render

everything. Each agent is empowered by an MLLM backbone and can access existing memory and call external tools when needed.

### 3.1. Strategist

The *Strategist* agent interfaces with advertisers to ground requirements for banner creation. Advertisers typically provide brand guidelines and specifications (*e.g.*, logo, desired dimensions). This agent determines key banner objectives including mood, tone, target audience, and primary purpose. To maintain brand identity, the *Strategist* analyzes the provided logo, trimming transparent padding to optimize space utilization while preserving visual content. Upon collecting all requirements, the *Strategist* stores this information in memory and transfers control to the *Background Designer* agent for the next phase of banner generation.

### 3.2. Background Designer

The *Background Designer* generates prompts for text-to-image (T2I) tools to create the visual canvas for banner ads. Implemented as a ReAct agent [71], it harmonizes advertiser requests, logo characteristics, and campaign objectives stored in memory to produce appropriate background visuals. The agent utilizes three specialized tools: *FindImagePath* to check for existing background images provided by advertisers, T2I to generate new visuals, and *TextChecker* to check if the generated image contains text. For new image generation, the agent employs a self-refinement loop [48, 70] to ensure text-free backgrounds, as text elements in the background layer would complicate the *Foreground Designer*'s work and potentially introduce difficult-to-remove gibberish text rendered directly into pixels. When text is detected by the *TextChecker* tool, the agent revises its prompt to discourage text generation and repeats the process for up to five iterations, as illustrated in Fig. 4 where a “Black Friday” background was successfully made text-free. During image generation, the agent selects the closest supported aspect ratio from the limited set available in the T2I tool that exceeds the target dimensions—ensuring higher initial resolution—and subsequently downscales to the target banner dimensions. This approach preserves visual fidelity while



Self-revised prompt: “A bold and eye-catching background with a contemporary design ~~for a Black Friday and Cyber Monday sale ...~~”

Figure 4. Prompt refinement for text-free backdrop creation.

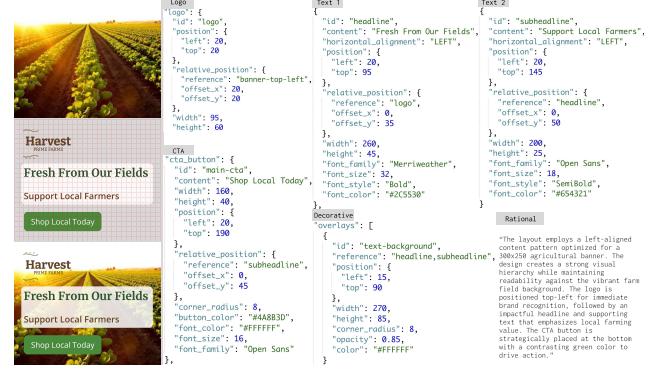


Figure 5. Element-wise blueprint as *Foreground Designer* output.

ensuring exact sizing for various display contexts. The resulting text-free background serves as an appropriate visual foundation, reserving space for the *Foreground Designer* to add text, CTA buttons, logos, and decorative elements in subsequent stages.

### 3.3. Foreground Designer

The *Foreground Designer* agent serves as the creative cornerstone of our system, generating a structured blueprint for all foreground elements encoded as a JSON-structured schema—defining precise positioning, styling, and content for logos, campaign texts, call-to-action buttons, and decorative elements. To maintain good alignment and overlap, we use relative positioning with reference properties to model the spatial relationship among banner components. An example of this element-wise blueprint from *Foreground Designer* is shown in Fig. 5. By transforming design decisions into structured data rather than rasterized pixels, the *Foreground Designer* creates a foundation for both visual coherence and complete element-level editability.

**Memory-augmented iterative design refinement.** The evolution of designs through iterative external feedback and refinement is a staple procedure in professional design [2]. Simulating this practice [48, 70], our memory-augmented iterative design refinement process, illustrated in Fig. 6, enhances banner ad designs by introducing an external critic *Design Reviewer*. At initial creation ( $t = 0$ ), the *Foreground Designer* generates the first blueprint based solely on the creative brief (banner request, background image, logo). The *Developer* renders this blueprint, and the *Design Reviewer* provides initial feedback, which is summarized and stored in memory. For the first refinement ( $t = 1$ ), the *Foreground Designer* refines the blueprint based on this feedback, comparing the new blueprint with the previous one and summarizing into a modification list (*i.e.*, “Compare & Conclude”). For subsequent refinements ( $t > 1$ ), both designer and reviewer access the memory that is con-

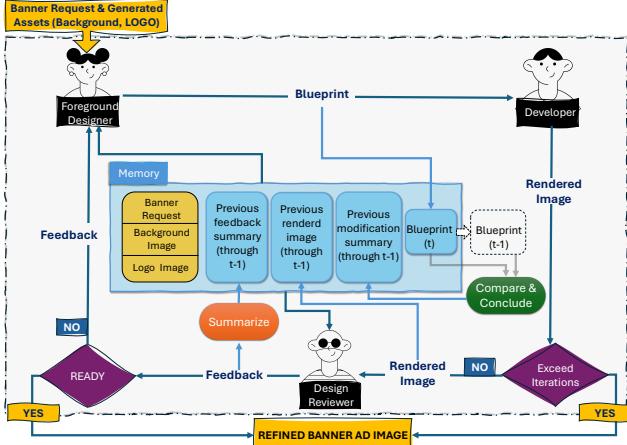


Figure 6. Memory-augmented iterative design refinement process.

tinuously augmented with new feedback, renderings, and modification summaries after each iteration, enabling increasingly informed decisions. This memory-augmented approach, where each iteration enriches the shared memory with new design artifacts and critiques, ensures both designer and reviewer maintain awareness of design evolution while continuously aligning with brand objectives. The process concludes when iterations are exceeded or the design is deemed production-ready by the reviewer, resulting in the final refined banner ad image.

### 3.4. Developer

The *Developer* plays a crucial role in converting the *Foreground Designer*'s blueprint into a banner ad image. We propose two implementations for the *Developer*: generating SVG code or Figma plugin code. Examples of codes in two formats are included in supplementary.

**SVG Code Generation.** Scalable Vector Graphics (SVG) is a widely-supported format for rendering two-dimensional graphics in web browsers and design software <sup>†</sup>. It is a declarative XML-based markup language with elements that directly define visual properties through coordinates, dimensions, and styling attributes. This approach provides flexibility and compatibility, as SVG is an open standard supported by various design tools. However, it requires an additional step of importing the SVG code into a design editing software such as Adobe Illustrator or Inkscape, and offers a more static representation of the design.

**Figma Plugin Code Generation.** Figma is a popular cloud-based design and prototyping tool used by designers worldwide <sup>†</sup>. By translating the blueprint into JavaScript

<sup>†</sup><https://www.w3.org/TR/SVG11/>

<sup>†</sup><https://www.figma.com/plugin-docs/>



(a) Working professionals.

(b) Kids.

Figure 7. Same campaign for different groups of audience.

code that adheres to the Figma Plugin API <sup>†</sup>, the *Developer* enables seamless integration with the Figma ecosystem. The generated code includes imperative programming instructions to programmatically create and manipulate nodes in the Figma document hierarchy. This approach builds the design through a sequence of operations rather than a static description. The advantage of this approach is that the banner ad image can be automatically rendered within Figma, eliminating the need for manual importing. This streamlines the design workflow and allows for quick previews and iterations while leveraging Figma's native component system and interactive editing capabilities.

## 4. BannerRequest400 Benchmark

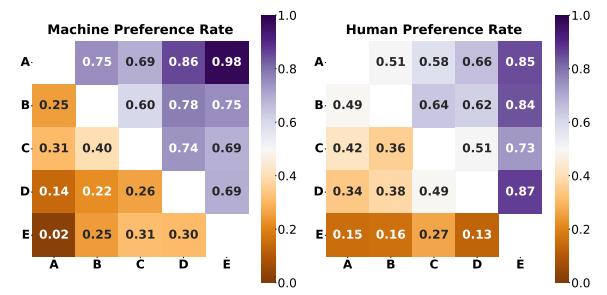
Existing design datasets, such as DESIGNERINTENTION [31], are limited to uni-modal designer intentions in text format and lack specific addressing of banner advertisement requirements. To address this gap, we introduce BannerRequest400, the first multimodal benchmark specifically designed for evaluating advertising banner generation systems. This novel benchmark pairs brand logos (visual modality) with corresponding banner design requests (textual modality). We curated 100 advertising design intentions from DESIGNERINTENTION and synthesized corresponding vector logos using Claude3.5 Sonnet [4], followed by expert review and refinement to visual fidelity, avoid bias, and maintain authentic logo aesthetics. Our empirical analysis, illustrated in Fig. 7, demonstrates how identical dental clinic promotion requests produce substantially different designs based on target audience—with professional-oriented banners emphasizing efficiency and results, while child-focused versions feature playful elements and incentives. Leveraging this insight, we employed GPT-4o[52] to systematically expand each intention into four distinct banner requests, each targeting a unique audience-purpose combination, resulting in 400 diverse specifications. To ensure industry relevance, we further extended these 400 requests across 13 standard banner dimensions, yielding a comprehensive evaluation set of 5,200 multimodal banner

<sup>†</sup><https://www.figma.com/plugin-docs/>

Method	TAA	LPS	CTAE	CPYQ	BIS	AQS
<i>Baselines</i>						
DALL-E3*	3.17	1.78	4.18	2.48	3.06	3.10
FLUX.1-schnell*	3.36	3.67	4.28	2.64	3.09	3.12
OpenCOLE	3.56	1.20	2.85	3.20	2.84	3.29
<i>BannerAgency</i>						
<i>SVG Implementation</i>						
[E] GPT-4o w/o BG	4.15	4.51	4.04	4.25	3.48	3.19
[D] Claude3.5-Sonnet w/o BG	4.40	<b>4.57</b>	4.63	4.48	4.26	3.77
[A] Claude3.5-Sonnet	<b>4.56</b>	4.43	<b>4.94</b>	<b>4.53</b>	<b>4.36</b>	<b>3.92</b>
<i>Figma Implementation</i>						
[C] GPT-4o	<b>4.39</b>	4.22	4.66	4.00	4.14	3.69
[B] Claude3.5-Sonnet	4.34	<b>4.33</b>	<b>4.81</b>	<b>4.19</b>	<b>4.22</b>	<b>3.85</b>

(a) Banner quality metrics (GPT-4o, human-validated in Tab. 2).

Table 1. Quantitative comparison on BannerRequest400 benchmark. \* denotes methods requiring explicit banner descriptions.



(b) System preference rates (evaluated by both GPT-4o and humans).

specifications. For benchmarking text-to-image generation methods, we further enhanced each request with detailed specifications that incorporate logo-specific visual characteristics using GPT-4o. BannerRequest400 enables the first rigorous evaluation of banner generation approaches across diverse design requests, facilitating advancement in automatic advertising banner design generation.

## 5. Experiments

### 5.1. Experimental Setup

**Algorithms.** For baseline approaches on our BannerRequest400 benchmark, we compare with two powerful pixel-based image generation models including closed-sourced DALL-E3 [6] and open-sourced FLUX.1-schnell [7] which is also our backbone for the T2I tool. We also compare with OpenCOLE [31, 33], a prompt-based graphic design model that relies on fine-tuning. For the backbone of our proposed training-free BannerAgency, we explore two multimodal LLM variants including GPT-4o (2024-10-21) [52] and Claude3.5-Sonnet (2024-10-22) [4] with temperature set as 0.3. We provide more details including all agent profile definitions and tool definitions in the supplementary material.

**Metrics.** To evaluate banner design quality, we developed six metrics (Tab. 2) from extensive literature survey on banner design principles [34, 36, 37, 46, 50, 55]. These metrics capture advertising-specific aspects overlooked by general design metrics [23, 31]: audience alignment, brand integration, call-to-action effectiveness, message clarity, brand consistency, and visual appeal—addressing both marketing functionality and aesthetic requirements critical for digital advertising. For completeness and comparison with prior metrics, we also include results using general design metrics in the supplementary material.

**Human study.** We conducted three human evaluations to assess different aspects of BannerAgency. First, a system preference study with 20 participants comparing banner designs generated by five BannerAgency variants across 20

Metric	Criteria	Pearson	Spearman	ICC
TAA	Measures how well the generated banner ad aligns with the given request, including the theme, target audience, and primary purpose.	0.854	0.840	0.922
LPS	Evaluates whether the logo is well-integrated into the design in terms of visibility, size, and positioning.	0.955	0.960	0.986
CTAE	Evaluates whether the Call-to-Action (CTA) is clear, engaging, and visually emphasized.	0.927	0.946	0.947
CPYQ	Evaluates the effectiveness of the headline, subheadline, and any other text in the banner ad, focusing on clarity, readability, persuasiveness, and grammatical correctness.	0.856	0.840	0.962
BIS	Measures how well the banner ad visually and stylistically aligns with the brand's identity beyond just logo placement. This includes color consistency, typography, imagery, and overall brand feel.	0.877	0.871	0.935
AQS	Measures the visual appeal, including color harmony, layout balance, typography, and overall design quality.	0.892	0.895	0.973

Table 2. **Banner design evaluation metrics and human-LLM correlation.** Each metric (rated 1-5) assesses a distinct aspect of banner design quality. Pearson, Spearman, and ICC correlation coefficients between human and GPT-4o ratings show strong agreement (all  $p < 0.001$ ). TAA: Target Audience Alignment; LPS: Logo Placement Score; CTAE: CTA Effectiveness; CPYQ: Copywriting Quality; BIS: Brand Identity Score; AQS: Aesthetic Quality Score. Detailed rating criteria are provided in supplementary.

randomly selected requests through pairwise comparisons. Second, a refinement effectiveness evaluation where 15 participants assessed quality progression across four iterations of 20 different banner designs, rating both initial and best versions on a 5-point scale. Third, a human-LLM correlation study validating the alignment between GPT-4o’s automated scoring and human perception across six proposed metrics, with 17-19 participants independently grading 25 representative images per metric. See appendix for details.

### 5.2. Results and Analysis

**Comparison with baselines.** We present quantitative and visual comparisons in Tab. 1(a) and Fig. 8, respectively. BannerAgency outperforms all baselines across all aspects. Baseline limitations include: fixed-size generation preventing non-standard dimensions like  $300 \times 250$ , failure to incorporate provided logos, and incorrect text rendering requiring correction [58]. Among BannerAgency variants, our preference heatmap in Tab. 1(b) shows Claude3.5-Sonnet surpasses GPT-4o ([D] vs. [E] and [B] vs. [C]), especially

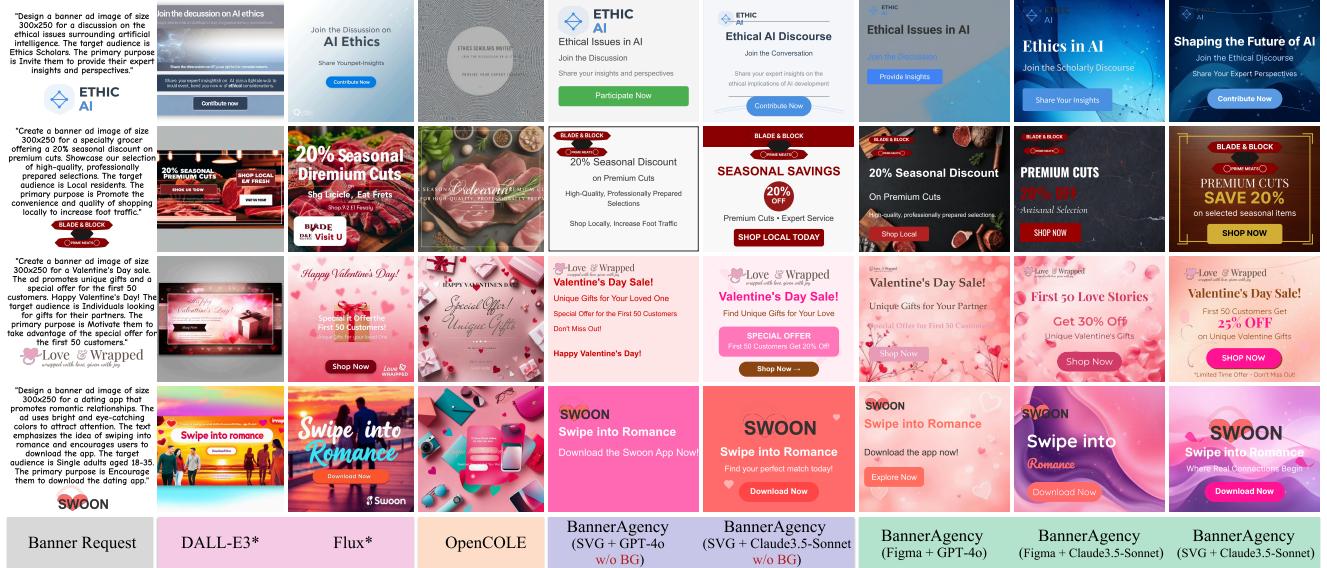


Figure 8. **Visual comparison on BannerRequest400 benchmark.** \* denotes methods requiring explicit banner descriptions as input (versus abstract requests) due to limited reasoning capabilities, whose detailed prompts are provided in the supplementary.

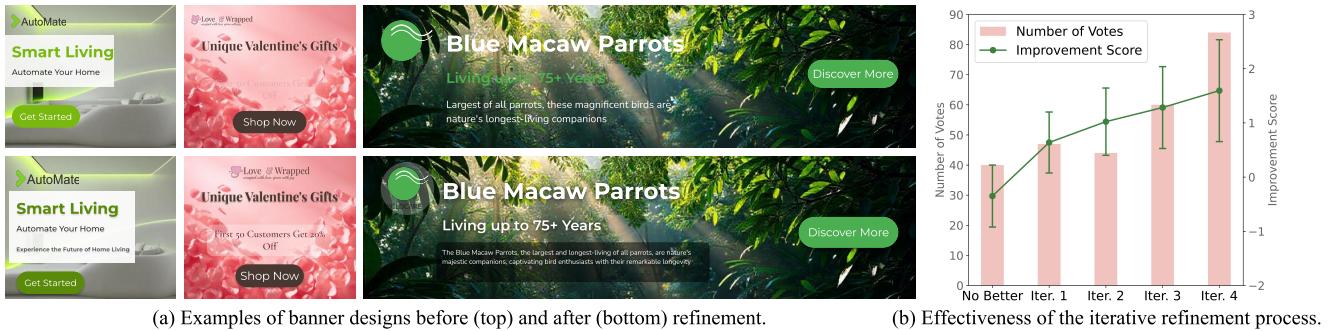


Figure 9. **BannerAgency’s design refinement.** Human study reveals later iterations received more votes (pink bars) with fourth iteration most preferred. Green line shows mean improvement between initial and selected versions, demonstrating significant quality gains.

in SVG implementation. Human evaluators showed no preference between SVG and Figma implementations ([A] vs. [B]). Notably, removing the Background Designer component (*i.e.*, w/o BG) significantly impairs visual appeal, resulting in plain and unengaging designs, particularly with the GPT-4o backbone. This underscores the critical role of specialized background generation in creating visually compelling banner advertisements.

**Refinement process.** Fig. 9(a) demonstrates the design improvement after the iterative refinement process. The initial designs are upgraded by enhancing content clarity, layout composition, and visual styling without radical redesigns. As shown in Fig. 9(b), our evaluation reveals a clear upward trend in quality, with improvement scores steadily increasing across iterations and later versions receiving more votes. Notably, the vote distribution indicates

that refinement effectiveness is design-context dependent—while the final iteration was most frequently preferred, many participants favored earlier iterations or the initial design. This finding highlights the value of maintaining a complete version track, allowing advertisers to select the specific iteration that best aligns with their vision, regardless of where it falls in the refinement sequence. This process also reflects how our component-based approach enables practical, editable design alternatives.

**Versatility to user prompts.** After presenting versatility to various sizes in Fig. 1, we showcase in Fig. 10 two more examples highlighting cross-genre aesthetic adaptation and cross-cultural linguistic flexibility. Fig. 10a presents six aesthetic styles for the same automation company request. We observe how the design adapt coherently to each style while maintaining color schemes that complement the Au-

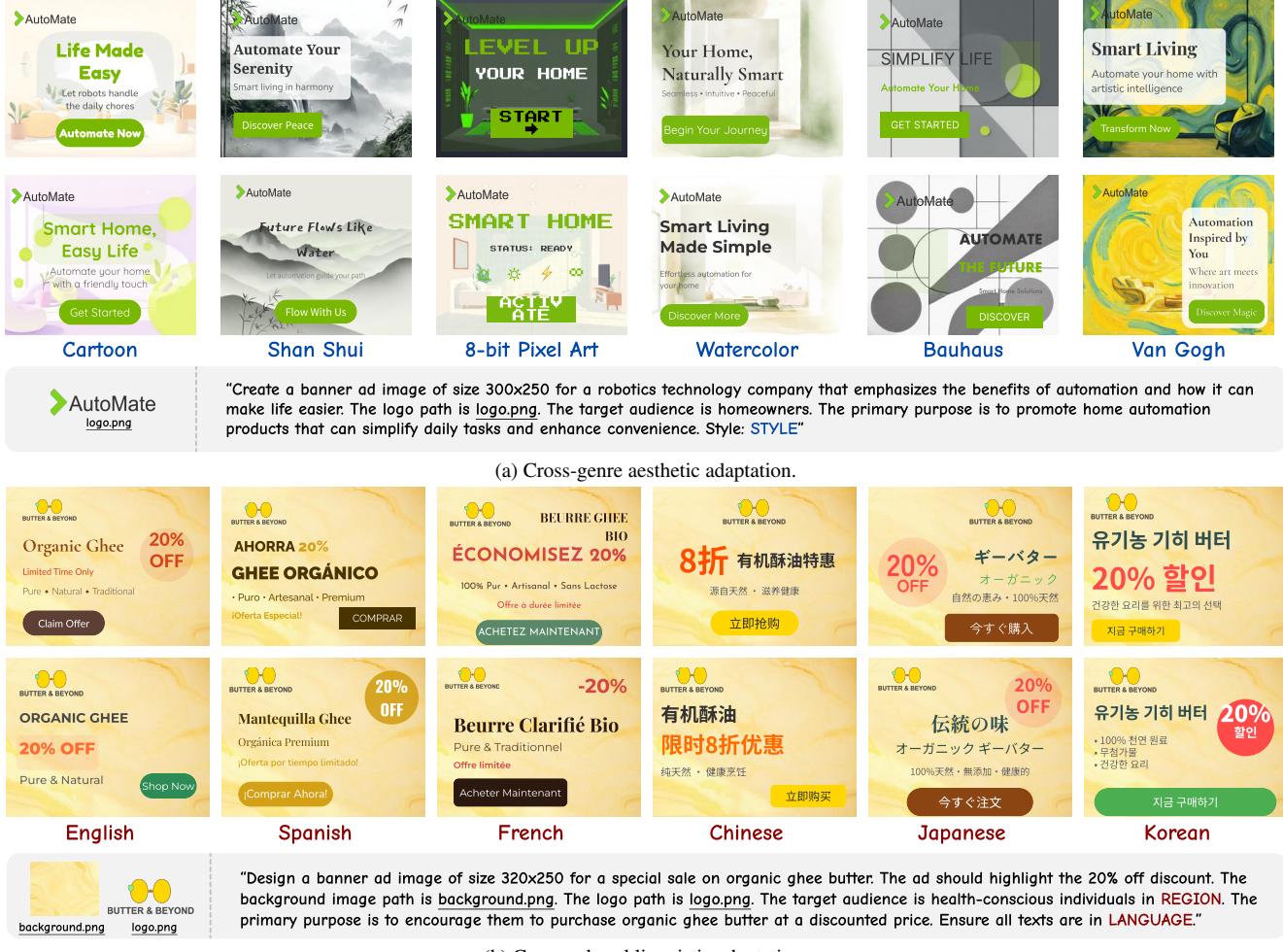


Figure 10. **BannerAgency’s design versatility.** We demonstrate the high versatility through (a) cross-genre aesthetic adaptation, and (b) cultural adaptation across diverse market contexts, all without requiring additional training or fine-tuning.

toMate logo. In Fig. 10b, we show how the system successfully maintains the 20% discount promotion across all versions while accommodating language-specific conventions and cultural preferences, which directly addresses the industry challenge of creating culturally appropriate marketing materials without specialized design teams. Empowered by the world knowledge of multimodal LLMs, our training-free BannerAgency demonstrates remarkable adaptability to diverse user requirements without any additional fine-tuning or domain-specific datasets.

## 6. Conclusion and Future Work

We introduced BannerAgency, a training-free MLLM agent system that transforms advertising banner design by simulating professional design teams through specialized agents (*Strategist*, *Background Designer*, *Foreground Designer*, and *Developer*). Our approach generates component-based,

editable designs in industry-standard formats from multimodal inputs without domain-specific fine-tuning, overcoming the limitations of pixel-based methods. Through our BannerRequest400 benchmark featuring 100 unique logos paired with 400 diverse banner requests, we demonstrate BannerAgency’s effectiveness in producing high-quality, versatile, and editable designs that kickstart the creative design process. This represents a significant step toward democratizing professional-grade design capabilities across the digital advertising landscape. In the future, we aim to extend BannerAgency by: (1) expanding multimodal input to incorporate product images, (2) introducing more sophisticated decorative elements beyond basic geometric shapes, 3) allowing for more hierarchical architecture for better collaboration among agents, and (4) considering conversion metric optimization to directly connect design choices with advertising performance.

## References

- [1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 3
- [2] ADCreative.ai. The best kept secrets about the iterative design process, 2024. Accessed: 2025-02-28. 4
- [3] Alibaba. The evolution of luban in designing one billion images, 2020. Accessed: 2025-02-28. 2
- [4] Anthropic. Claude 3.5 sonnet, 2024. 2, 5, 6
- [5] Bannerwise. Does size matter? understanding ad sizes and when to use them, 2022. Accessed: 2025-02-28. 2
- [6] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *OpenAI*, 2023. 2, 3, 6
- [7] BlackForestLab. Announcing black forest labs, 2024. 3, 6
- [8] Canva. Free banner maker, 2025. Accessed: 2025-02-28. 2
- [9] Chao-Ting Chen and Hen-Hsen Huang. Integrating llm, vlm, and text-to-image models for enhanced information graphics: A methodology for accurate and visually engaging visualizations. In *IJCAI*, pages 8627–8630, 2024. 3
- [10] Jin Chen, Ju Xu, Gangwei Jiang, Tiezheng Ge, Zhiqiang Zhang, Defu Lian, and Kai Zheng. Automated creative optimization for e-commerce advertising. In *WWW*, pages 2304–2313, 2021. 2
- [11] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser-2: Unleashing the power of language models for text rendering. In *ECCV*, pages 386–402. Springer, 2024. 2
- [12] Xingye Chen, Wei Feng, Zhenbang Du, Weizhen Wang, Yanyin Chen, Haohan Wang, Linkai Liu, Yaoyu Li, Jinyuan Zhao, Yu Li, et al. Ctr-driven advertising image generation with multimodal large language models. In *WWW*, 2025. 2
- [13] Yutao Cheng, Zhao Zhang, Maoke Yang, Hui Nie, Chunyuan Li, Xinglong Wu, and Jie Shao. Graphic design with large multimodal model. *arXiv preprint arXiv:2404.14368*, 2024. 3
- [14] Saemi Choi, Kiyoharu Aizawa, and Nicu Sebe. Fontmatcher: font image paring for harmonious digital graphic design. In *IUI*, pages 37–41, 2018. 2
- [15] The LinkedIn community. How do you know if your graphics are working in your marketing campaigns?, 2025. Accessed: 2025-02-28. 2
- [16] Zhenbang Du, Wei Feng, Haohan Wang, Yaoyu Li, Jingsen Wang, Jian Li, Zheng Zhang, Jingjing Lv, Xin Zhu, Junsheng Jin, et al. Towards reliable advertising image generation using human feedback. In *ECCV*, pages 399–415. Springer, 2024. 2
- [17] Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, et al. Agent ai: Surveying the horizons of multimodal interaction. *arXiv preprint arXiv:2401.03568*, 2024. 2, 3
- [18] Adobe Express. Design graphic banners for free, 2025. Accessed: 2025-02-28. 2
- [19] Bi Qi Fong and John See. Branddiffusion: Multimodal personalized marketing visual content generation. In *ACM MM Workshop on Multimedia Content Generation and Evaluation: New Methods and Practice*, pages 72–77, 2024. 2
- [20] Jiaxin Ge, Zora Zhiruo Wang, Xuhui Zhou, Yi-Hao Peng, Sanjay Subramanian, Qinyue Tan, Maarten Sap, Alane Suhr, Daniel Fried, Graham Neubig, et al. Autopresent: Designing structured visuals from scratch. *arXiv preprint arXiv:2501.00912*, 2025. 3
- [21] Shunan Guo, Zhuochen Jin, Fuling Sun, Jingwen Li, Zhaorui Li, Yang Shi, and Nan Cao. Vinci: an intelligent graphic design system for generating advertising posters. In *CHI*, pages 1–17, 2021. 2
- [22] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *CVPR*, pages 14953–14962, 2023. 2
- [23] 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*, New York, NY, USA, 2024. Association for Computing Machinery. 6, 16, 17
- [24] Yingqing He, Zhaoyang Liu, Jingye Chen, Zeyue Tian, Hongyu Liu, Xiaowei Chi, Runtao Liu, Ruibin Yuan, Yazhou Xing, Wenhui Wang, et al. Llms meet multimodal generation and editing: A survey. *arXiv preprint arXiv:2405.19334*, 2024. 3
- [25] Hao Hu, Chao Zhang, and Yanxue Liang. A study on the automatic generation of banner layouts. *Computers & Electrical Engineering*, 93:107269, 2021. 2
- [26] Ziniu Hu, Ahmet Iscen, Aashi Jain, Thomas Kipf, Yisong Yue, David A Ross, Cordelia Schmid, and Alireza Fathi. Scenecraft: An llm agent for synthesizing 3d scenes as blender code. In *ICML*, 2024. 2
- [27] Qirui Huang, Min Lu, Joel Lanir, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Graphimind: Llm-centric interface for information graphics design. *arXiv preprint arXiv:2401.13245*, 2024. 2, 3
- [28] Shir Iluz, Yael Vinker, Amir Hertz, Daniel Berio, Daniel Cohen-Or, and Ariel Shamir. Word-as-image for semantic typography. *TOG*, 42(4):1–11, 2023. 2
- [29] Naoto Inoue, Kotaro Kikuchi, Edgar Simo-Serra, Mayu Otani, and Kota Yamaguchi. Towards flexible multi-modal document models. In *CVPR*, pages 14287–14296, 2023. 3
- [30] Naoto Inoue, Kotaro Kikuchi, Edgar Simo-Serra, Mayu Otani, and Kota Yamaguchi. Layoutdm: Discrete diffusion model for controllable layout generation. In *CVPR*, pages 10167–10176, 2023. 3
- [31] Naoto Inoue, Kento Masui, Wataru Shimoda, and Kota Yamaguchi. Opencole: Towards reproducible automatic graphic design generation. In *CVPR Workshops*, pages 8131–8135, 2024. 2, 3, 5, 6, 16, 17
- [32] Shoma Iwai, Atsuki Osanai, Shunsuke Kitada, and Shinichiro Omachi. Layout-corrector: Alleviating layout sticking phenomenon in discrete diffusion model. In *ECCV*, pages 92–110. Springer, 2024. 3
- [33] Peidong Jia, Chenxuan Li, Yuhui Yuan, Zeyu Liu, Yichao Shen, Bohan Chen, Xingru Chen, Yinglin Zheng, Dong

- Chen, Ji Li, et al. Cole: A hierarchical generation framework for multi-layered and editable graphic design. *arXiv preprint arXiv:2311.16974*, 2023. 2, 3, 6
- [34] Robert Katai. 5 powerful banner ad a/b testing ideas you should try, 2024. Accessed: 2025-02-28. 6
- [35] Kotaro Kikuchi, Naoto Inoue, Mayu Otani, Edgar Simo-Serra, and Kota Yamaguchi. Multimodal markup document models for graphic design completion. *arXiv preprint arXiv:2409.19051*, 2024. 3
- [36] Cassandra King. Essential practices for high-converting banner ad designs, 2020. Accessed: 2025-02-28. 6
- [37] Robin Landa. *Advertising by design: generating and designing creative ideas across media*. John Wiley & Sons, 2016. 2, 6
- [38] Hsin-Ying Lee, Lu Jiang, Irfan Essa, Phuong B Le, Haifeng Gong, Ming-Hsuan Yang, and Weilong Yang. Neural design network: Graphic layout generation with constraints. In *ECCV*, pages 491–506. Springer, 2020. 3
- [39] Guandong Li and Xian Yang. Smartbanner: intelligent banner design framework that strikes a balance between creative freedom and design rules. *Multimedia Tools and Applications*, 82(12):18653–18667, 2023. 2
- [40] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742. PMLR, 2023. 3
- [41] Jiawei Lin, Jiaqi Guo, Shizhao Sun, Zijiang Yang, Jian-Guang Lou, and Dongmei Zhang. Layoutprompter: awaken the design ability of large language models. In *NeurIPS*, pages 43852–43879, 2023. 2, 3
- [42] Jinpeng Lin, Min Zhou, Ye Ma, Yifan Gao, Chenxi Fei, Yangjian Chen, Zhang Yu, and Tiezheng Ge. Autoposter: A highly automatic and content-aware design system for advertising poster generation. In *ACM MM*, pages 1250–1260, 2023. 3
- [43] Jieru Lin, Danqing Huang, Tiejun Zhao, Dechen Zhan, and Chin-Yew Lin. Spot the error: Non-autoregressive graphic layout generation with wireframe locator. In *AAAI*, pages 3413–3421, 2024. 3
- [44] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, pages 34892–34916, 2023. 2, 3
- [45] Zeyu Liu, Weicong Liang, Zhanhao Liang, Chong Luo, Ji Li, Gao Huang, and Yuhui Yuan. Glyph-byt5: A customized text encoder for accurate visual text rendering. In *ECCV*, pages 361–377. Springer, 2024. 2
- [46] Ritu Lohtia, Naveen Donthu, and Edmund K Hershberger. The impact of content and design elements on banner advertising click-through rates. *Journal of Advertising Research*, 43(4):410–418, 2003. 6
- [47] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024. 2
- [48] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dzirri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. In *NeurIPS*, pages 46534–46594, 2023. 4
- [49] Paridhi Maheshwari, Nitish Bansal, Surya Dwivedi, Rohan Kumar, Pranav Manerikar, and Balaji Vasan Srinivasan. Exemplar based experience transfer. In *IUI*, pages 673–680, 2019. 2
- [50] Smridhi Malhotra. Essential practices for high-converting banner ad designs, 2024. Accessed: 2025-02-28. 6
- [51] ManyPixels. The complete graphic design pricing guide (2025 update), 2025. Accessed: 2025-02-28. 12
- [52] OpenAI. Hello gpt-4o, 2024. 2, 5, 6
- [53] Peter O’Donovan, Aseem Agarwala, and Aaron Hertzmann. Learning layouts for single-pagegraphic designs. *TVCG*, 20(8):1200–1213, 2014. 3
- [54] Seonmi Park, Inhwan Bae, Seunghyun Shin, and Hae-Gon Jeon. Kinetic typography diffusion model. In *ECCV*, pages 166–185. Springer, 2024. 2
- [55] Helen Robinson, Anna Wysocka, and Chris Hand. Internet advertising effectiveness: the effect of design on click-through rates for banner ads. *International journal of advertising*, 26(4):527–541, 2007. 2, 6
- [56] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 3
- [57] Mohammad Amin Shabani, Zhaowen Wang, Difan Liu, Nanxuan Zhao, Jimei Yang, and Yasutaka Furukawa. Visual layout composer: Image-vector dual diffusion model for design layout generation. In *CVPR*, pages 9222–9231, 2024. 3
- [58] Wataru Shimoda, Naoto Inoue, Daichi Haraguchi, Hayato Mitani, Seichi Uchida, and Kota Yamaguchi. Type-r: Automatically retouching typos for text-to-image generation. *arXiv preprint arXiv:2411.18159*, 2024. 2, 6
- [59] Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing. In *ICLR*, 2024. 2
- [60] Praneetha Vaddamanu, Vinay Aggarwal, Bhanu Prakash Reddy Guda, Balaji Vasan Srinivasan, and Niyati Chhaya. Harmonized banner creation from multimodal design assets. In *CHI Extended Abstracts*, pages 1–7, 2022. 2
- [61] Sreekanth Vempati, Korah T Malayil, V Sruthi, and R Sandeep. Enabling hyper-personalisation: Automated ad creative generation and ranking for fashion e-commerce. In *Fashion Recommender Systems*, pages 25–48. Springer, 2020. 2
- [62] Yael Vinker, Tamar Rott Shaham, Kristine Zheng, Alex Zhao, Judith E Fan, and Antonio Torralba. Sketchagent: Language-driven sequential sketch generation. *arXiv preprint arXiv:2411.17673*, 2024. 3
- [63] Shiya Wang, Qi Liu, Yicheng Zhong, Zhilong Zhou, Tiezheng Ge, Defu Lian, and Yuning Jiang. Creagan: An automatic creative generation framework for display advertising. In *ACM MM*, pages 7261–7269, 2022. 2

- [64] Shaodong Wang, Yunyang Ge, Liuhan Chen, Haiyang Zhou, Qian Wang, Xinhua Cheng, and Li Yuan. Prompt2poster: Automatically artistic chinese poster creation from prompt only. In *ACM MM*, pages 10716–10724, 2024. 2, 3
- [65] Zhenyu Wang, Aoxue Li, Zhenguo Li, and Xihui Liu. Genartist: Multimodal llm as an agent for unified image generation and editing. In *NeurIPS*, pages 128374–128395, 2025. 2, 3
- [66] Haohan Weng, Danqing Huang, Yu Qiao, Zheng Hu, Chin-Yew Lin, Tong Zhang, and CL Chen. Designen: A pipeline for controllable design template generation. In *CVPR*, pages 12721–12732, 2024. 3
- [67] Junlin Xie, Zhihong Chen, Ruifei Zhang, Xiang Wan, and Guanbin Li. Large multimodal agents: A survey. *arXiv preprint arXiv:2402.15116*, 2024. 2, 3
- [68] Hao Yang, Jianxin Yuan, Shuai Yang, Linhe Xu, Shuo Yuan, and Yifan Zeng. A new creative generation pipeline for click-through rate with stable diffusion model. In *WWW*, pages 180–189, 2024. 2
- [69] Xuyong Yang, Tao Mei, Ying-Qing Xu, Yong Rui, and Shipeng Li. Automatic generation of visual-textual presentation layout. *TOMM*, 12(2):1–22, 2016. 2
- [70] Zhengyuan Yang, Jianfeng Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. Idea2img: Iterative self-refinement with gpt-4v (ision) for automatic image design and generation. In *ECCV*, 2024. 3, 4
- [71] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *ICLR*, 2023. 4
- [72] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 3
- [73] Yunke Zhang, Kangkang Hu, Peiran Ren, Changyuan Yang, Weiwei Xu, and Xian-Sheng Hua. Layout style modeling for automating banner design. In *ACM MM Workshops*, pages 451–459, 2017. 2
- [74] Zhen Zhao, Jingqun Tang, Binghong Wu, Chunhui Lin, Shu Wei, Hao Liu, Xin Tan, Can Huang, Yuan Xie, et al. Harmonizing visual text comprehension and generation. In *NeurIPS*, 2024. 2
- [75] Hao Zheng, Xinyan Guan, Hao Kong, Jia Zheng, Hongyu Lin, Yaojie Lu, Ben He, Xianpei Han, and Le Sun. Pptagent: Generating and evaluating presentations beyond text-to-slides. *arXiv preprint arXiv:2501.03936*, 2025. 3, 16
- [76] Xinru Zheng, Xiaotian Qiao, Ying Cao, and Rynson WH Lau. Content-aware generative modeling of graphic design layouts. *TOG*, 38(4):1–15, 2019. 3
- [77] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *ICLR*, 2024. 3

# BannerAgency: Advertising Banner Design with Multimodal LLM Agents

## Supplementary Material

In this appendix, we present more details on our experimental setup (Sec. 7), human study setup (Sec. 8), text-free background generation (Sec. 9), and foreground variation generation (Sec. 9). We also show detailed statistics results on the refinement human study (Sec. 10) and general graphic design evaluation (Sec. 10). The detailed prompt for text-to-image generation models in Fig. 8 of our main paper is presented in Sec. 11. Lastly, we showcase more results from our BannerAgency in Sec. 12.

### 7. More Details on Experimental Setup

**Cost analysis.** We present a detailed cost breakdown in USD for generating banner advertisements of size  $300 \times 250$  pixels in Tab. 3. Our BannerAgency upscales images to  $1024 \times 864$  pixels to maintain visual quality before rendering. Costs are calculated based on Claude 3.5 Sonnet API pricing (\$3/MTok for input, \$15/MTok for output) and FLUX-1 [schnell] image generation (\$0.0027/MP). We compare both Figma and SVG implementation approaches across three scenarios: single image generation, generation with four-rounds of iterative refinement, and multi-variation generation with four design alternatives as detailed in Sec. 9. The dramatic difference between the two implementations stems from fundamental implementation constraints: while SVG can be generated directly as code that LLMs frequently encounter in training data, Figma plugin JavaScript is uncommon in LLM training corpora. To ensure reliable execution, our Figma implementation requires providing templates where the Developer agent only needs to call predefined functions, significantly increasing token usage. Notably, both implementation approaches offer extraordinary cost-effectiveness compared to hiring professional designers, who typically charge \$50-150 per hour [51]. Even our most expensive scenario (Figma with feedback at \$1.213) represents less than 2.5% of traditional design costs, offering transformative economics for advertising campaigns requiring multiple banner variations.

Implementation	Scenario	Prompt Tokens	Completion Tokens	LLM Cost (\$)	Image Cost (\$)	Total Cost (\$)	Cost/Image (\$)
Single image	Figma	78,266	5,470	0.317	0.002	0.319	0.319
	SVG	15,921	2,672	0.088	0.002	0.090	0.090
With refinement	Figma	284,677	23,772	1.211	0.002	1.213	1.213
	SVG	54,362	12,663	0.353	0.002	0.355	0.355
Four variations	Figma	261,150	15,213	1.012	0.008	1.020	0.255
	SVG	24,865	6,006	0.165	0.008	0.173	0.043

Table 3. BannerAgency System Cost Breakdown

**Figma vs. SVG format** Both implementation approaches have their merits, depending on the user’s preferences and

the design tools they are comfortable with. The SVG code generation offers broad compatibility and flexibility, while the Figma plugin code generation provides a more integrated and automated experience within the Figma ecosystem. The code examples are provided in Fig. 11.

**Agent tool definition.** We document the tool list in Tab. 4. The Background Designer agent creates visual backdrops for banner advertisements that align with brand identity and campaign objectives. It leverages existing assets when available and employs a refinement process to ensure text-free backgrounds through detection and regeneration cycles. The system handles dimensional requirements by adapting to supported aspect ratios and resizing outputs to match target specifications. The Developer Agent transforms design specifications into editable implementations through a systematic process. It loads and customizes Figma plugin templates, manages asset references, compiles implementation code, and handles rendering by interfacing with Figma’s API. Throughout this workflow, comprehensive archiving preserves complete implementations alongside their assets, enabling advertisers to receive fully editable component-based designs rather than static images.

**Metrics rating criteria.** We present the detailed rating criteria for each metric in Tab. 5.

### 8. More Details on Human Study

We present screenshots of the survey instructions in Fig. 12 to demonstrate how we structured each human evaluation.

**Preference.** The preference experiment evaluated user preferences across different variants of the banner ad generation system. It consisted of 20 banner ad requests, each implemented with 5 different system variants, resulting in 10 possible pairwise comparisons per request. Twenty participants completed the same set of comparisons, viewing pairs of banner ads generated from identical requests but using different system variants. For each pair, participants were asked to select which banner ad they preferred. The 20 banner requests were randomly selected to match the distribution of the full 400-request dataset, ensuring that the selected subset was representative of the overall performance characteristics of each variant. To avoid potential bias, the order of presentation for the banner pairs was randomized for each participant.

```

1  async function createText(content, initialFontSize,
2    maxWidth, maxHeight, fontFamily = "Inter", fontStyle = "Regular",
3    color = { r: 0, g: 0.5, b: 0 }, alignment = 'CENTER') {
4    const text = figma.createText();
5    await safeLoadFonts(text, fontFamily, fontStyle);
6    text.characters = content;
7    text.fills = [{ type: 'SOLID', color }];
8    text.textAlignHorizontal = alignment;
9
10   let currentFontSize = initialFontSize;
11   text.fontSize = currentFontSize
12   text.resize(maxWidth, maxHeight);
13   text.textAutoResize = "HEIGHT";
14   while ((text.width > maxWidth || text.height > maxHeight) && currentFontSize > 1){
15     currentFontSize--;
16     text.fontSize=currentFontSize
17   }
18   return text;
19 }
20
21 // Create headline text
22 const headline = await createText(
23   "Your Time, Your Smile",
24   36,
25   180,
26   88,
27   "Raleway",
28   "ExtraBold",
29   { r: 10/255, g: 36/255, b: 99/255 },
30   "LEFT"
31 );
32 headline.x = 16;
33 headline.y = 70;
34 headline.name = "headline";
35 frame.appendChild(headline);

```

(a) Figma Plugin Code Snippet

```

1  <svg width="728" height="90" xmlns="http://www.w3.org/2000/svg">
2    <!-- Background Image -->
3    <image href="data:image/png;base64,[ENCODED]" width="728" height="90" preserveAspectRatio="xMidYMid slice"/>
4
5    <!-- Semi-transparent overlay -->
6    <rect x="0" y="0" width="728" height="90" fill="black" opacity="0.3"/>
7
8    <!-- Logo -->
9    <image href="data:image/png;base64,[ENCODED]" x="20" y="15" width="86" height="50" preserveAspectRatio="xMidYMid"/>
10
11   <!-- Decorative wine glass illustration -->
12   <path d="M680 25 Q685 15, 690 25 L695 60 Q685 65, 675 60 Z" fill="none" stroke="rgba(255,255,255,0.7)" stroke-width="1"/>
13
14   <!-- Main Text -->
15   <text x="140" y="35" font-family="Playfair Display" font-size="24" fill="white">Elevate Your Wine Experience</text>
16
17   <!-- Discount Text -->
18   <text x="140" y="60" font-family="Inter" font-size="20" fill="#FFD700">35% OFF Premium Wine Glass Sets</text>
19
20   <!-- CTA Button -->
21   <g>
22     <rect x="550" y="25" width="160" height="40" rx="20" fill="#FFD700"/>
23     <text x="580" y="50" font-family="Inter" font-weight="600" font-size="16" fill="#B80000">SHOP NOW</text>
24   </g>
25
26   <!-- Decorative Elements -->
27   <line x1="120" y1="15" x2="120" y2="75" stroke="rgba(255,255,255,0.3)" stroke-width="1"/>
28   <circle cx="120" cy="45" r="3" fill="#FFD700"/>
29 </svg>

```

(b) SVG Code Snippet

Figure 11. Figma vs. SVG code format.

Tool Name	Description
update_image_list	Updates the imageList constant in a Figma plugin's UI HTML file with a provided list of image filenames, enabling dynamic referencing of assets within the plugin interface.
read_plugin_template	Reads and validates a Figma plugin code template file, ensuring it's a proper JavaScript file containing essential Figma plugin components before using it as a foundation for customization.
save_plugin_code	Saves the finalized JavaScript code to the code.js file in the specified Figma plugin working directory after validating that the directory contains a manifest.json file.
render_and_save_image	Executes the Figma plugin by running an AppleScript, retrieves the latest node ID, and uses the Figma API to export and save the rendered banner design as a PNG image at the specified path.
backup_plugin_folder	Creates a backup copy of the entire plugin folder with all assets (including background image, logo, and rendered output) to preserve the complete state of the implementation for future reference or versioning.
create_unique_image_name	Generates a unique filename by combining a timestamp with the banner name to ensure distinct identification of each banner design output (e.g., "20240328_154227_kids.dentist.clinic.banner").
T2I	Generates banner backgrounds from textual descriptions by interfacing with a text-to-image model, adapting dimensions to supported aspect ratios, and resizing outputs to match target banner dimensions. Includes negative prompting to avoid text generation in backgrounds.
FindImagePath	Analyzes user input to locate existing image files that could serve as backgrounds, identifies valid image paths with common extensions (.jpg, .png, etc.), filters out logo images, verifies file existence, and resizes found images to match target banner dimensions.
TextChecker	Evaluates generated or existing background images to detect the presence of any text elements, using a multimodal LLM to analyze the visual content and determine if the background is suitable or requires regeneration.

Table 4. List of tool definitions.

**Refinement.** The refinement experiment evaluated the effectiveness of an iterative refinement approach for banner ad image generation. It involved 20 sets of banner designs,

with each set containing four iterations of refinement (modifying only foreground elements while keeping backgrounds consistent). Fifteen participants reviewed all 20 sets and

<b>Abbr.</b>	<b>Metric Name</b>	<b>Definition</b>	<b>Rating Criteria</b>
TAA	Target Audience Alignment	Measures how well the generated banner ad aligns with the given request, including the theme, target audience, and primary purpose.	<p><b>5</b> – Perfectly aligns with the request (theme, audience, purpose are all clearly reflected).</p> <p><b>4</b> – Mostly aligns, but minor details could be improved.</p> <p><b>3</b> – Somewhat aligns, but key elements are missing or unclear.</p> <p><b>2</b> – Barely aligns, with major missing or incorrect elements.</p> <p><b>1</b> – Does not align with the request at all.</p>
LPS	Logo Score	Evaluates whether the logo is well-integrated into the design in terms of visibility, size, and positioning.	<p><b>5</b> – Logo is well-placed, clearly visible, proportionate, and blends seamlessly.</p> <p><b>4</b> – Logo is well-placed but could be slightly improved (e.g., minor size or position adjustments).</p> <p><b>3</b> – Logo is visible but not ideally placed (e.g., too small, too large, or slightly obstructed).</p> <p><b>2</b> – Logo placement is poor (e.g., difficult to notice, awkward positioning).</p> <p><b>1</b> – Logo is either missing or completely misplaced.</p>
AQS	Aesthetic Quality Score	Measures the visual appeal, including color harmony, layout balance, typography, and overall design quality.	<p><b>5</b> – Visually outstanding, professional design, well-balanced, with harmonious colors and readable text.</p> <p><b>4</b> – Well-designed, but small refinements could enhance it.</p> <p><b>3</b> – Acceptable but has notable design flaws (e.g., poor contrast, unbalanced elements).</p> <p><b>2</b> – Visually weak, with noticeable design mistakes.</p> <p><b>1</b> – Poor design, lacks professionalism or coherence.</p>
CTAE	Call-to-Action Effectiveness	Evaluates whether the Call-to-Action (CTA) is clear, engaging, and visually emphasized.	<p><b>5</b> – CTA is clear, compelling, well-placed, and visually prominent.</p> <p><b>4</b> – CTA is effective but could be slightly improved (e.g., contrast, size).</p> <p><b>3</b> – CTA is present but lacks emphasis or clarity.</p> <p><b>2</b> – CTA is weak, hard to notice, or poorly worded.</p> <p><b>1</b> – No clear CTA is present.</p>
CPYQ	Copywriting Quality	Evaluates the effectiveness of the headline, subheadline, and any other text in the banner ad, focusing on clarity, readability, persuasiveness, and grammatical correctness.	<p><b>5</b> – Copy is clear, engaging, grammatically correct, and persuasive, making the message effective.</p> <p><b>4</b> – Copy is well-written but could be slightly improved (e.g., minor word choice refinements).</p> <p><b>3</b> – Copy is somewhat effective but has issues in clarity, grammar, or persuasiveness.</p> <p><b>2</b> – Copy is weak, hard to read, contains noticeable grammatical mistakes, or lacks impact.</p> <p><b>1</b> – Copy is unclear, irrelevant, or difficult to read due to poor design or bad wording.</p>
BIS	Brand Identity Score	Measures how well the banner ad visually and stylistically aligns with the brand's identity beyond just logo placement. This includes color consistency, typography, imagery, and overall brand feel.	<p><b>5</b> – Strong brand consistency; the banner design aligns well with the provided logo and conveys a recognizable brand identity.</p> <p><b>4</b> – Mostly aligns, but minor refinements could improve brand consistency.</p> <p><b>3</b> – Somewhat aligns, but noticeable inconsistencies exist (e.g., off-brand colors, incorrect typography).</p> <p><b>2</b> – Weak brand alignment, only the logo represents the brand while other design choices feel unrelated.</p> <p><b>1</b> – No brand identity is reflected; the banner appears generic or disconnected from the brand.</p>

Table 5. Banner design rating criteria with a 5-point scale for each metric.

performed two tasks: (1) selecting which iteration among the four achieved the best refinement performance, and (2) rating both the initial version and their chosen best version

on a scale of 1 to 5. This experiment aimed to determine whether the refinement process consistently improved designs and at which iteration the optimal results were typi-

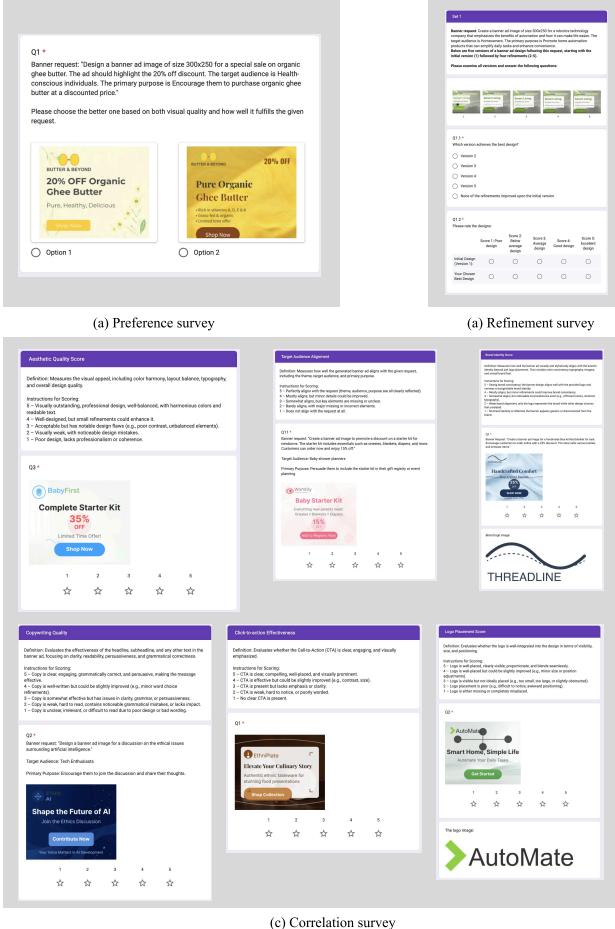


Figure 12. The survey examples.

cally achieved.

**Correlation.** The correlation experiment assessed whether six proposed metrics for banner ad quality (TAA: Target Audience Alignment, LPS: Logo Placement Score, CTAE: CTA Effectiveness, CPYQ: Copywriting Quality, BIS: Brand Identity Score, and AQS: Aesthetic Quality Score) aligned with human perception. For each metric and each point on a 1-to-5 scale, five images previously graded by GPT-4o at that score level were selected, resulting in 25 images per metric and 150 images total. Ten participants graded these images, with the presentation order shuffled to prevent bias. Each participant evaluated all images across the six metrics to determine whether human evaluators would assign similar scores to those given by GPT-4o.

## 9. More Details on Design Generation

**Details of Text-free Background Generation.** Algorithm 1 shows the detailed behavior of the text-free background generation.

### Algorithm 1 Text-free Background Design Agent Behavior

**Input:** User requirements  $R$ , Logo characteristics  $L$

**Output:** Background image path  $P$ , Description  $D$  (if generated)

**Agent Tools:**

FINDIMAGEPATH :  $R \mapsto (P, \text{bool})$  Primary search tool

T2I :  $D \mapsto P$  Image generation tool

TEXTCHECKER :  $P \mapsto \text{bool}$  Verification tool

```

1:  $P, \text{path\_exists} \leftarrow \text{FINDIMAGEPATH}(R)$ 
2: if  $\text{path\_exists}$  then
3:   return  $P$ 
4: else
5:    $\text{attempts} \leftarrow 0$ 
6:    $\text{contains\_text} \leftarrow \text{TRUE}$ 
7:   while  $\text{contains\_text}$  AND  $\text{attempts} < 5$  do
8:     Analyze  $R$  and  $L$  to formulate background description  $D$ 
9:      $P \leftarrow \text{T2I}(D)$ 
10:     $\text{contains\_text} \leftarrow \text{TEXTCHECKER}(P)$ 
11:     $\text{attempts} \leftarrow \text{attempts} + 1$ 
12:    if  $\text{contains\_text}$  then
13:      Refine  $D$  to eliminate text-producing elements
14:    end if
15:   end while
16:   return  $P, D$ 
17: end if

```



Figure 13. Variation.

**Memory-aware Design Variation.** Our memory-aware approach generates diverse alternatives by allowing the *Foreground Designer* to reference previous designs. Each variation  $V_i$  is conditioned on:

$$V_i = f(B, L, A, \{V_1, V_2, \dots, V_{i-1}\}), \quad (1)$$

where  $B$  represents background image properties,  $L$  denotes logo characteristics,  $A$  encapsulates advertising objectives, and  $\{V_1, V_2, \dots, V_{i-1}\}$  are previously generated designs. By integrating these multimodal constraints, each new variation positions itself distinctly while preserving brand consistency and advertising goals. This method balances visual coherence with design diversity, exploring variations in meaningful dimensions such as layout, color

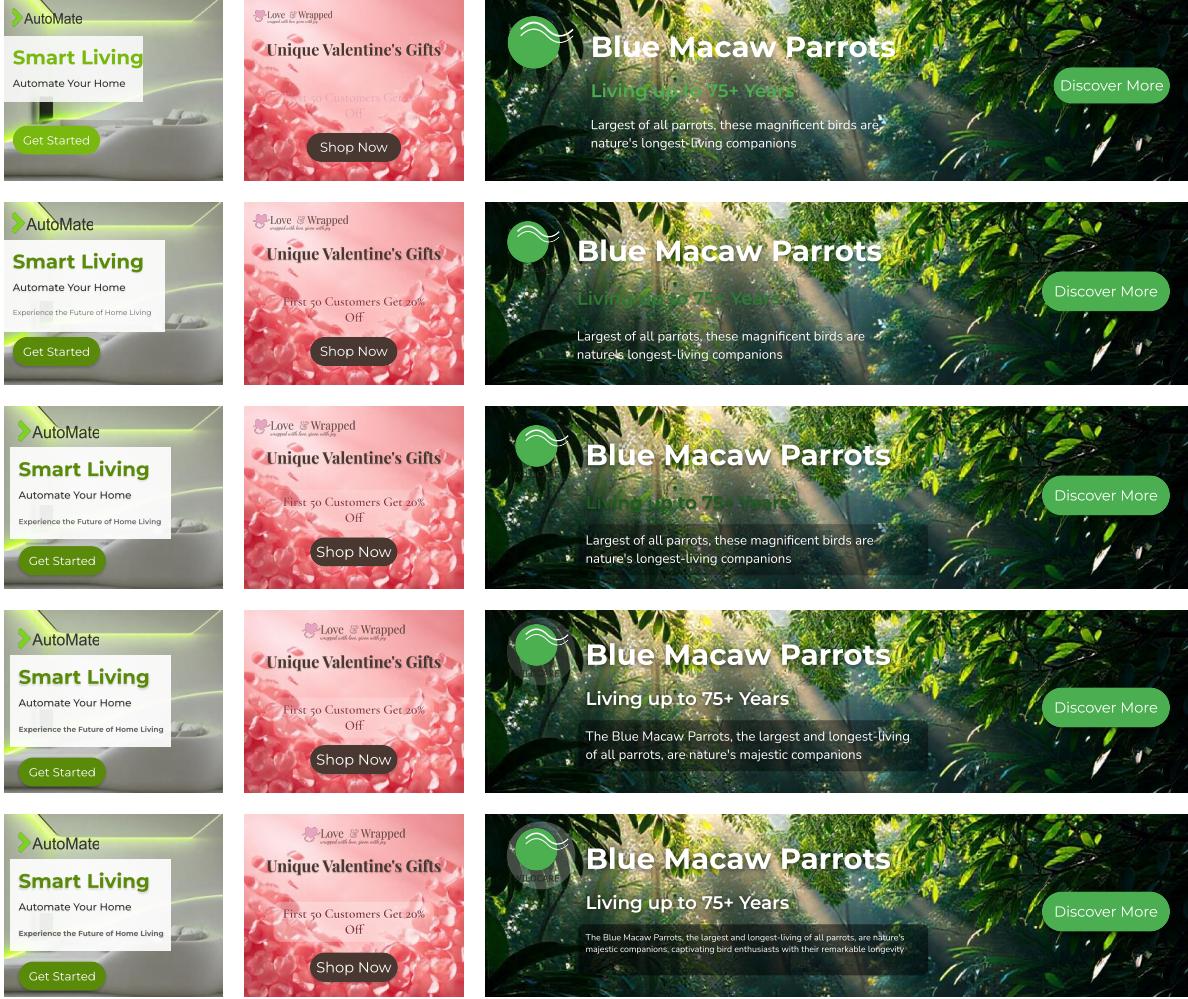


Figure 14. BannerAgency’s design refinement. From top to bottom shows design evolution at each refinement step.

scheme, copywriting, and typography while maintaining alignment with the original creative brief. Fig. 13 showcases four different designs generated with our memory-aware design variation generation mechanism. We can observe the distinct difference between each new variant and its previous designs. The resulting variations offer advertisers multiple creative interpretations of the same brief, each with distinct visual strategies while maintaining brand consistency.

## 10. More Results Analysis

**Detailed refinement performance.** To complement Fig.9(a) in our main paper, we present the complete design evolution in Fig. 14. We also present the full statistics from human study on the refinement process in Tab. 6. We observe a statistically significant improvement in quality ratings, increasing from 2.56 ( $\sigma = 0.89$ ) for initial designs to 3.55 ( $\sigma = 0.77$ ) for refined versions ( $p < 0.001$ ).

Step	#votes	Improvement scores	std
No better	40	-0.350	0.572
Iter. 1	47	0.638	0.562
Iter. 2	44	1.023	0.621
Iter. 3	60	1.283	0.755
Iter. 4	84	1.595	0.940

Table 6. Full statistics of refinement human study

**Evaluation with general graphic design metrics.** We compute general graphic design metrics [23, 31] in Tab. 7 to provide complementary analysis. Following PPTA-agent [75], we compute the FID (Fréchet Inception Distance) score between the generated banner ad designs (400) and exemplar designs (3355)<sup>†</sup> using 64-dimensional feature vectors. We observe that Figma implementations gener-

<sup>†</sup><https://huggingface.co/datasets/PeterBrendan/AdImageNet>

Method	FID $\downarrow$	DL	CR	TC	GI	INV	ALM	OLP	WS
<i>Baselines</i>									
DALL-E3	2.65	7.18	7.21	5.87	7.58	6.13	4.69	4.42	4.32
Flux	3.22	7.24	6.92	5.91	7.67	6.02	4.70	4.46	4.23
OpenCOLE	3.99	7.34	8.11	6.47	7.92	6.70	4.87	4.57	4.53
<i>BannerAgency</i>									
<i>SVG Implementation</i>									
GPT-4o - BG	17.76	6.75	7.73	5.98	5.84	4.82	5.07	<b>6.44</b>	5.26
Claude3.5-Sonnet - BG	13.61	7.22	8.18	6.56	6.36	5.52	6.01	5.91	<b>5.51</b>
Claude3.5-Sonnet	<b>4.25</b>	<b>7.47</b>	<b>8.43</b>	<b>6.92</b>	<b>7.14</b>	<b>6.19</b>	<b>5.77</b>	5.40	5.08
<i>Figma Implementation</i>									
GPT-4o	<b>2.60</b>	7.22	8.17	6.40	7.47	6.30	4.73	4.85	4.77
Claude3.5-Sonnet	4.99	<b>7.46</b>	<b>8.40</b>	<b>6.84</b>	<b>7.47</b>	<b>6.39</b>	<b>5.16</b>	<b>5.20</b>	<b>5.17</b>

Table 7. Evaluation metrics for generic design quality. Primary metrics include: Design and Layout (DL), Content Relevance (CR), Typography and Color (TC), Graphics and Images (GI), and Innovation (INV) [31], as well as technical aspects [23] such as Alignment (ALM), Overlap prevention (OLP), and White Space utilization (WS). All metrics are evaluated on a 10-point scale by GPT-4o, except for FID (Fréchet Inception Distance), which compares the similarity between generated designs and exemplar ones.

ally achieve lower FID scores (GPT-4o: 2.60, Claude3.5-Sonnet: 4.99) than SVG implementations, suggesting closer visual similarity to real-world examples. The evaluation metrics, all adopted from prior work, include GPT-4o assessment of five primary aspects: (i) design and layout (DL), (ii) content relevance (CR), (iii) typography and color (TC), (iv) graphics and images (GI), and (v) innovation (INV) [31], alongside technical metrics on alignment (ALM), overlap prevention (OLP), and white space utilization (WS) [23]. The results reveal that BannerAgency implementations are competitive with or surpass baseline methods across most metrics, particularly in content relevance (CR) where Claude3.5-Sonnet achieves 8.43, exceeding all baselines. The SVG implementation with Claude3.5-Sonnet notably outperforms in white space utilization (5.51) compared to baselines (highest: 4.53 with OpenCOLE). These results are presented in the supplementary materials as these generic design metrics, while informative, don't fully capture banner advertisement-specific requirements such as attention guidance, call-to-action effectiveness, and brand consistency, which are more central to our paper's focus.

## 11. Detailed Prompt in Fig.8 of Main Paper

- “Please create a 300x250 banner ad image with a background that features a subtle, sophisticated gradient of light blue and white, evoking a sense of calm and intellectual engagement. Overlay this with a faint, abstract pattern of interconnected nodes and lines, symbolizing the complex network of AI and ethical considerations.”

*Centered at the top, include the text ‘Join the Discussion on AI Ethics’ in a bold, dark gray sans-serif font. Below this, in a slightly smaller font, add ‘Share Your Expert Insights’ in the same blue color as our logo. At the bottom center, place a prominent CTA button in blue [#4285F4] with white text that reads ‘Contribute Now’. Position our logo in the bottom left corner, ensuring it is clearly visible but not overpowering the main message.”*

- “Create a 300x250 banner ad image with a background showcasing a vibrant, high-quality selection of premium cuts of meat, professionally prepared and arranged to look fresh and enticing. Use warm, inviting lighting to emphasize the quality and freshness of the products. Overlay the text ‘20% Seasonal Discount on Premium Cuts’ in bold, white font at the top center of the banner. Below this, include the text ‘Shop Local, Eat Fresh’ in a slightly smaller, italicized white font. Place a prominent, eye-catching CTA button at the bottom center of the banner with the text ‘Visit Us Today’ in bold, dark red [#C8102E] font on a white background. Position the Blade & Block logo in the bottom left corner, ensuring it is clearly visible but not overpowering the main message.”
- “Create a 300x250 banner ad image with a romantic Valentine’s Day theme. The background should feature a soft, dreamy blend of pink and red hues with subtle heart patterns. In the center, place an image of a beautifully wrapped gift box with a ribbon, symbolizing unique gifts. Overlay the text ‘Happy Valentine’s Day!’ in an elegant, large serif font at the top. Below that, in a slightly

*smaller but bold font, add ‘Special Offer for the First 50 Customers!’ and ‘Unique Gifts for Your Loved One’ to highlight the promotion. Place a prominent CTA button at the bottom center with the text ‘Shop Now’ in white on a dark brown [#4A2E2B] background. Position the Love & Wrapped logo in the bottom right corner, ensuring it is clearly visible but not overpowering the main message.”*

4. *“Please create a 300x250 banner ad image for a dating app targeting single adults aged 18-35. Use a bright and eye-catching color scheme with a background image that features a vibrant, romantic scene such as a sunset beach or a cityscape with couples walking hand-in-hand. Overlay the text ‘Swipe into Romance’ prominently in a bold, playful font, ensuring it stands out. Below this, include a CTA button with the text ‘Download Now’ in a contrasting color to grab attention. Place the Swoon logo in the bottom right corner, ensuring it is clearly visible but does not overpower the main message. The overall design should be modern, clean, and appealing to the target audience, encouraging them to download the app.”*

## 12. More Visual Results

**BannerAgency with Different sizes on BannerRequest400 benchmark.** To showcase the performance of our BannerAgency in different sizes of banner designs, we present more visual results from BannerAgency with different implementations in Fig. 15. Note how SVG implementation always includes decorative elements such as curves, lines, etc.

**More algorithm comparison.** Fig. 16 presents more visual comparisons among algorithms.

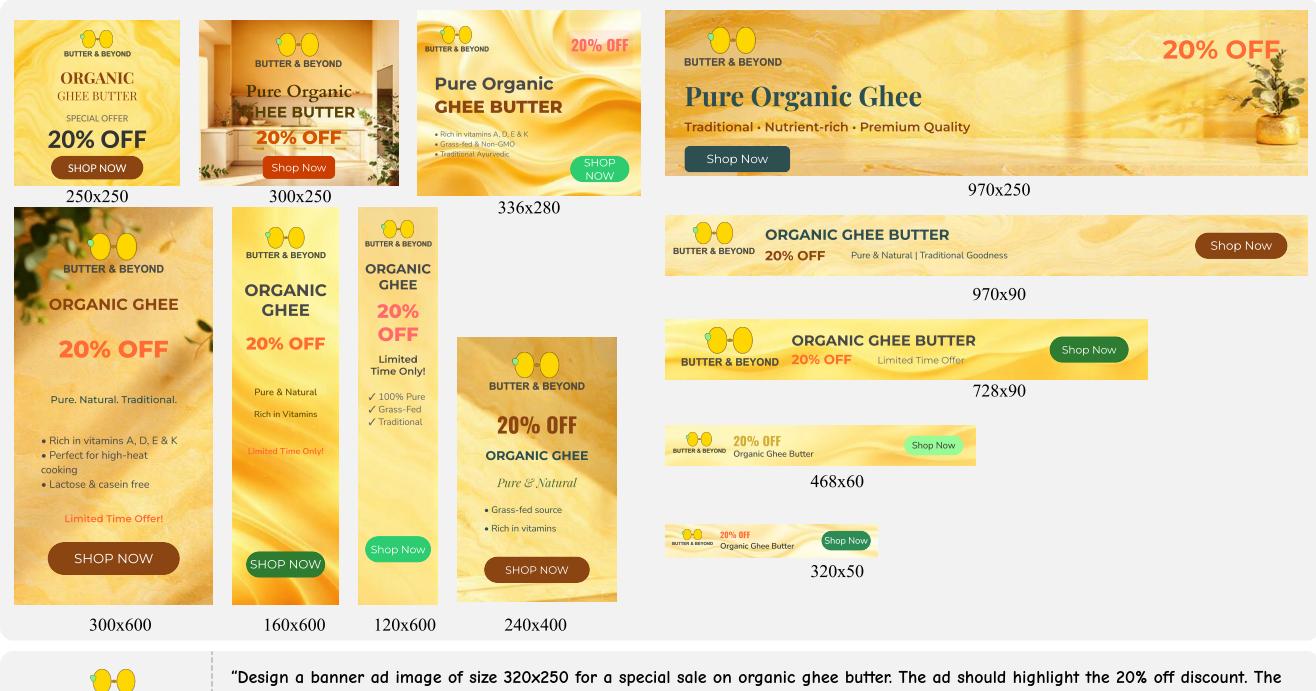
**More banner designs.** We present four more banner designs with the same advertiser but with different target audience and purposes in Fig. 17.



(a) BannerAgency with SVG implementation. Sizes: Skyscraper (160 × 600), leaderboard (728 × 90), medium rectangle (300 × 250).



(b) BannerAgency with Figma implementation. Sizes: Skyscraper (160 × 600), leaderboard (728 × 90), medium rectangle (300 × 250).



(c) Various sizes for same banner request.

Figure 15. More results from BannerAgency of different sizes.

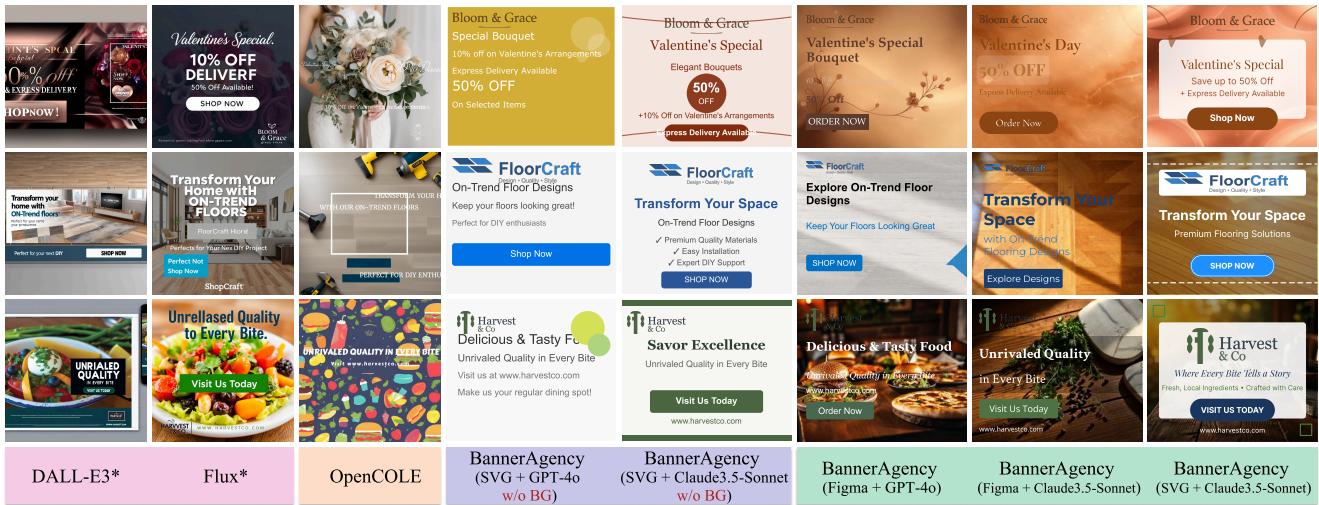
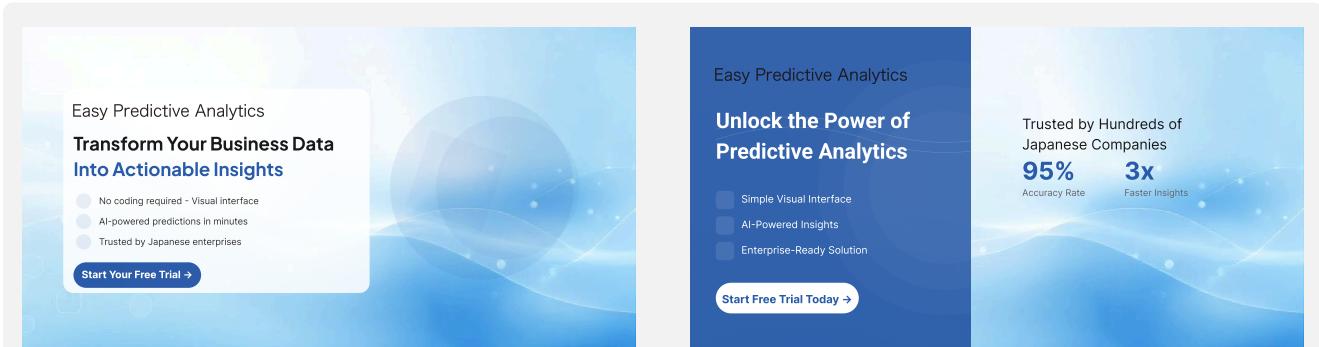
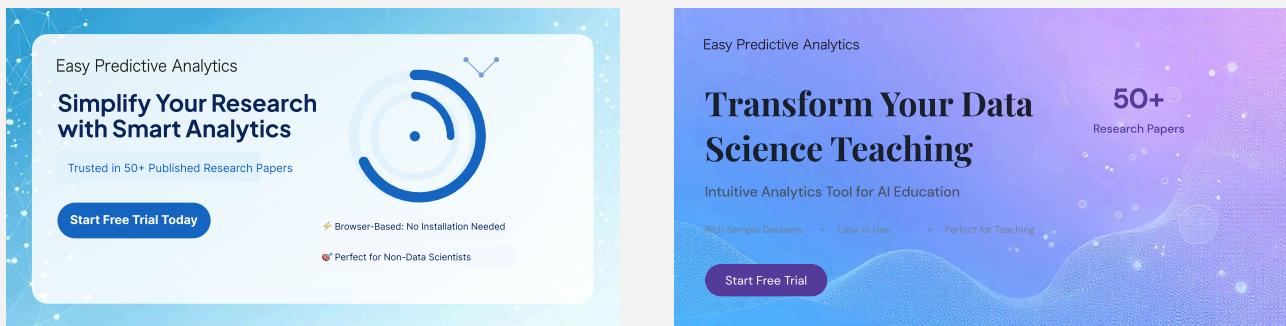


Figure 16. More visual comparisons of different algorithms.



"Design a banner ad image of size 1200x628 for Easy Predictive Analytics, a user-friendly, AI-powered SaaS tool for predictive analytics. It is trusted by hundreds of Japanese companies. The target audience is business professionals who want to introduce predictive analytics with ease. The primary purpose of the ad is to encourage them to visit the landing page to sign up for the free trial of Easy Predictive Analytics."



"Design a banner ad image of size 1200x628 for Easy Predictive Analytics, a simple yet powerful predictive analytics tool used in more than 50 research papers. It works directly in a web browser and helps non-data analytics experts to conduct predictive analytics with ease. The target audience is researchers who want to run predictive analytics on their data, but don't have enough time or data analytics skills. The primary purpose of the ad is to encourage the researchers to visit the landing page to sign up for the free trial of Easy Predictive Analytics."

"Design a banner ad image of size 1200x628 for Easy Predictive Analytics, a simple yet powerful predictive analytics tool used in more than 50 research papers. The tool offers abundant sample data and is intuitive to use, making it suitable for learning AI concepts in AI or data science classes. The target audience is teachers or instructors in charge of AI or data science classes. The primary purpose of the ad is to encourage them to visit the landing page to sign up for the free trial of Easy Predictive Analytics."

Figure 17. More example banner designs from BannerAgency with the same advertiser but different target audiences and purposes.