TeamFusion: Supporting Open-ended Teamwork with Multi-Agent Systems

Jiale Liu[1] **Victor S. Bursztyn**[2] **Lin Ai**[3] **Haoliang Wang**[2]
Sunav Choudhary[2] **Saayan Mitra**[2] **Qingyun Wu**[1,4]
[1]Pennsylvania State University [2]Adobe Research [3]Columbia University [4]AG2ai, Inc.
{jiale.liu, qingyun.wu}@psu.edu
lin.ai@cs.columbia.edu
{victor.bursztyn, haowang, schoudha, smitra}@adobe.com

Abstract

In open-ended domains, teams must reconcile diverse viewpoints to produce strong deliverables. Answer aggregation approaches commonly used in closed domains are ill-suited to this setting, as they tend to suppress minority perspectives rather than resolve underlying disagreements. We present TeamFusion, a multi-agent system designed to support teamwork in open-ended domains by: 1. Instantiating a proxy agent for each team member conditioned on their expressed preferences; 2. Conducting a structured discussion to surface agreements and disagreements; and 3. Synthesizing more consensus-oriented deliverables that feed into new iterations of discussion and refinement. We evaluate TeamFusion on two teamwork tasks where team members can assess how well their individual views are represented in team decisions and how consensually strong the final deliverables are, finding that it outperforms direct aggregation baselines across metrics, tasks, and team configurations.

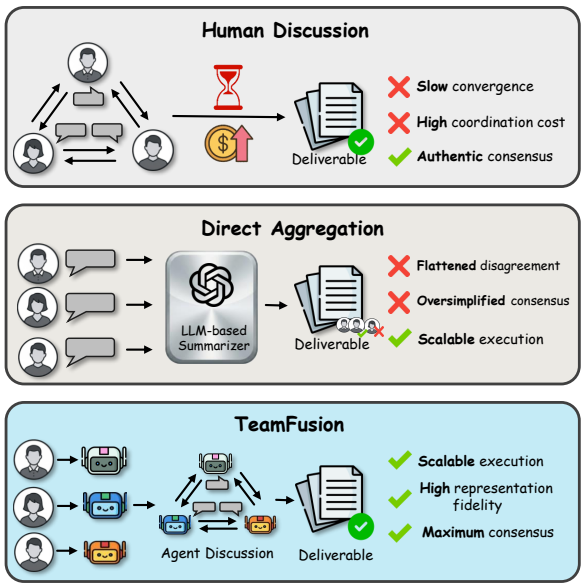

Figure 1: Illustration of TeamFusion versus baselines. While human discussion is slow and direct aggregation loses nuance, TeamFusion leverages agent-based discussion to combine fast execution with the high representation fidelity and maximum consensus.

1 Introduction

Many group decisions are open-ended: there is no single correct answer, but multiple plausible options that trade off values, constraints, and risk (Black, 1948; Kiesler and Sproull, 1992; Kraemer and King, 1988). In these settings, success is not "matching the gold label," but producing a deliverable that group participants recognize as reflecting their distinct preferences and rationales (Fisher, 1970). However, arriving at such a deliverable is expensive: teams must surface hidden assumptions, productively discuss disagreements, and negotiate acceptable trade-offs, which create communication bottlenecks and high costs at scale (Romney et al., 2025; Rogelberg et al., 2006).

Large language models (LLMs) appear promising for decision support because they can digest large amounts of text and draft deliverables that people can critique and revise (Zhang et al., 2025; Naveed et al., 2025). Yet many existing LLM usages in group settings still follow direct aggregation: concatenate inputs and generate a single recommendation (Bhaskar et al., 2023; Li et al., 2024, 2023), or collapse rationales into an "average" feedback (Zhu et al., 2025; Huang et al., 2023). This approach is ill-suited for open-ended teamwork for two reasons. First, a single-shot aggregate can be hard to audit and may introduce ungrounded claims, which is problematic when the deliverable must be attributable to participants' stated reasons (Huang et al., 2025; Parcalabescu and Frank, 2024; Liu et al., 2023; Yu et al., 2025). Second, direct aggregation can suppress disagreements instead of attempting to resolve them in a refined deliverable (Zhu et al., 2025; Laban et al., 2023; Zhang et al., 2024b; Wang et al., 2023). In other words, teamwork is not only about capturing shared com-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14435–14456

monsense, but also discussing local disagreements to make deliverables well-rounded overall.

This gap motivates our work: *how can we generate deliverables that preserve diverse individual viewpoints while helping teams to converge?* We argue that closing this gap requires modeling the iterative consensus-seeking process. In open-ended decisions, key information emerges when perspectives respond to one another: participants clarify opinions, challenge missing cases, and refine proposals in light of others' objections. A system that skips this step must implicitly guess the structure of disagreement from raw text, which is precisely where viewpoint erasure occurs.

We introduce TeamFusion, a general multi-agent framework for open-ended teamwork support. TeamFusion (i) instantiates a proxy agent for each team member, conditioned on their expressed preferences; (ii) runs a structured discussion to make agreements and disagreements explicit; and (iii) synthesizes the discussion into an editable deliverable that records trade-offs and supporting reasons. Our central hypothesis is that explicitly modeling team members and their interaction yields deliverables that are both more representative of diverse viewpoints and more useful for decision-making than direct aggregation.

We evaluate TeamFusion on two teamwork tasks where team members can judge how well their individual views are represented in team decisions and how consensually strong the final deliverables are. The results indicate that TeamFusion outperforms baselines across metrics, tasks, backbone models, and team configurations. Our contributions are:

1. We propose TeamFusion, a framework for open-ended decision support in teams. By modeling the process of consensus-seeking, TeamFusion synthesizes deliverables that cover wider viewpoints while driving convergence.
2. We propose a scalable, human-in-the-loop evaluation protocol for open-ended team tasks. By decoupling preference collection from interaction, our protocol overcomes the logistical bottlenecks of synchronous team studies, allowing for rigorous, large-scale evaluation of AI tools with professional domain experts.
3. Our results generalize across text and multimodal tasks, as well as teams of different sizes, showing that structured agent interaction yields higher-quality deliverables.

2 Related Work

2.1 Multi-agent Systems

Recent work has explored "societies" of LLM with agents that interact via structured dialogue (Piatti et al., 2024; Park et al., 2023). General-purpose orchestration frameworks such as AutoGen (Wu et al., 2024), MetaGPT (Hong et al., 2023) and LangChain (Chase, 2022) make it easier to construct multi-agent systems via role assignment, tool use, and customizable interaction protocols. Within this broader trend, multi-agent debate has emerged as a simple but effective recipe: multiple model instances propose answers, critique one another, refine a final response (Du et al., 2023; Chan et al., 2023; Liang et al., 2024). Extensive work has shown that by orchestrating and integrating agent responses, the system can generate outputs that are more factual (Du et al., 2023; Chern et al., 2024; Kim et al., 2024), creative (Liang et al., 2024; Hu et al., 2025), and functionally correct (Sun et al.; Zhang and Xiong, 2025; Song et al., 2024; Zhang et al., 2024a). Whereas debate frameworks primarily optimize for correctness or factuality, our focus is to use structured discussion to find and expand agreements, disagreements, and trade-offs in open-ended decisions.

2.2 LLMs for Group Consensus

Developing systems for group consensus has been a long-reaching question in NLP, with pre-LLM work building meeting corpora and identifying decision-related dialogue to support teams' shared understanding (Carletta et al., 2006; Shriberg et al., 2004; Orwig et al., 1997). With the advent and prevalence of LLMs, recent work increasingly leverages the model as a facilitator that steers deliberation: (Tessler et al., 2024) proposed "Habermas Machine," an LLM mediator that can help small groups find common ground in democratic deliberation, while structured conversational interventions can counter group think and improve how teams scrutinize AI advice during collective decisions (Chiang et al., 2024), and prompt-tuned mediation strategies can de-escalate or reframe online conflict toward agreement (Govers et al., 2024). Building on this emerging view of LLMs as facilitators, TeamFusion supports convergence of open-ended teamwork by representing each participant with a conditioned proxy agent, orchestrating a structured multi-party discussion, and synthesizing the discussion into a deliverable, iteratively.

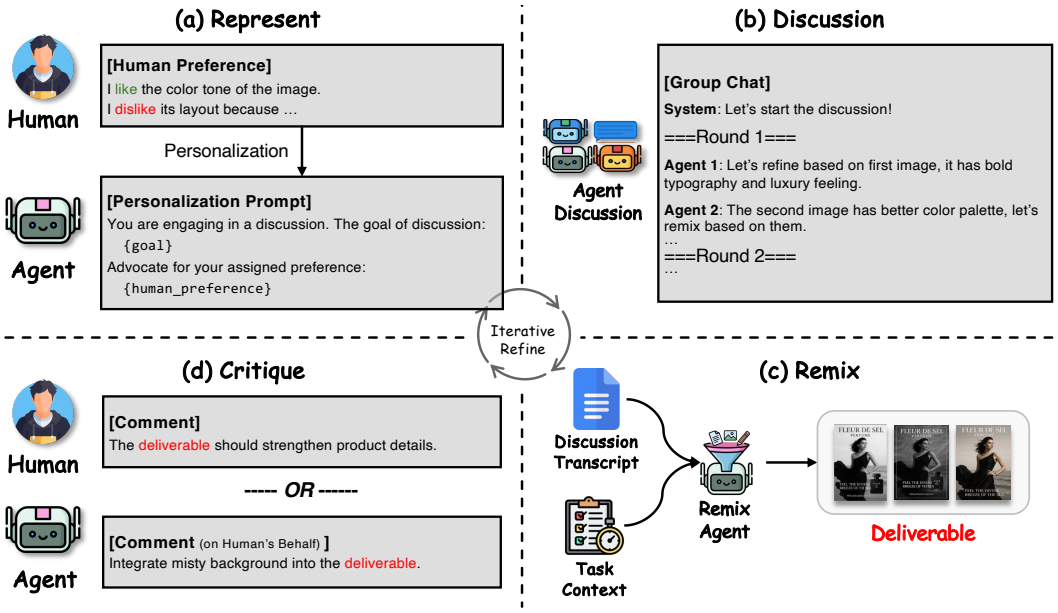

Figure 2: The overview of the TeamFusion framework. It consists of four phases: (1) Represent: We extract human preference labels as agents; (2) Discussion: The agents abstracted from human preference engage in a structured discussion; (3) Remix: The discussion transcript along with task context are remixed into a final deliverable used directly for downstream decision.; (4) Critique and Refine: The agent or human leave critiques based on generated deliverable, and the system iterates again on improving the deliverable.

3 TeamFusion Framework

Problem setup. We study *open-ended* teamwork, where the goal is to produce a deliverable that (i) preserves distinct viewpoints and constraints and (ii) helps a team move toward an acceptable outcome. Given a task context c and a set of N team members $\{u_1, \ldots, u_N\}$, each providing task-specific preference E_i, TeamFusion outputs a deliverable y intended to be directly usable.

Overview As shown in Figure 2, TeamFusion consists of four phases: (1) we instantiate one proxy agent per participant from their preferences; (2) proxy agents engage in a structured group discussion; (3) a remix phase converts the discussion into an editable deliverable, and (4) the system iterates this loop to refine the deliverable.

3.1 Represent

For each participant u_i, we create a proxy agent a_i designed to argue from u_i's perspective during the group discussion. Training a separate model per participant is impractical in realistic settings: per-user data are sparse, training is computationally expensive, and models would quickly become obsolete as preferences shift. Instead, we adopt an in-context personalization approach. Concretely, we encode the participant's evidence E_i into a structured system prompt π_i that specifies: (i) the agent's role and collaborative objective, (ii) domain and communication constraints, and (iii) the participant-specific preferences. Our goal is not to fully model a participant's identity, but to ensure the agent's contributions are recognizably aligned with that participant's expressed perspectives.

3.2 Discuss

Agent Roles One TeamFusion run consists of N participant proxy agents: $A = \{a_1, \ldots, a_N\}$. Proxy agents contribute proposals and critiques from their participant's perspective.

Conversation State The discussion proceeds in a shared group-chat environment. At step t, the controller maintains a message history $H_t = [m_1, \ldots, m_t]$, where each message $m_k = (name, content)$. Agents do not hold additional private state. At each turn, the controller selects a speaker $a \in A$ and prompts the underlying LLM with the agent's system prompt π_a and the current history H_t. This ensures that all agents reason over the same dialogue context.

Turn-Taking Protocol We adopt a simple but effective round-robin protocol inspired by the classic divergence–convergence model of creative processes (Acar and Runco, 2019; Runco and Acar,

2012) and nominal group technique (Dowling and St. Louis, 2000), giving each proxy agent a fixed number of speaking turns and cycling deterministically through agents to ensure equal opportunities to contribute. On a proxy turn, a_i receives H_t and is instructed to respond in light of its participant evidence and advance the discussion toward a recommendation. The discussion ends once all proxy agents exhaust their allotted turns.

3.3 Remix

After the debate concludes, TeamFusion converts the accumulated discussion into a final deliverable for the open-ended task at hand. A remixing agent takes as input the original task context c and the full discussion history H_T, and produces a deliverable that combines all reasoning over disagreements and points of convergence. The remixed deliverable is intended to be directly consumable by humans experts on the task. Implementation details for each task are provided in Appendix F.3.

3.4 Iterative Refinement

TeamFusion can be applied once to obtain a single deliverable, or used iteratively to gradually refine outputs. In the iterative setting, the deliverable from one round is treated as a new proposal that reenters the discussion: proxy agents are given access to the updated deliverable alongside the original context and asked to critique and build upon it in a subsequent discussion. This refinement loop allows the system to successively narrow in on options that better reflect surfaced preferences and rationales.

4 Task 1: Civic Comment Synthesis

We begin by evaluating TeamFusion in a civic decision-support setting, where a small group must turn diverse free-form public comments into a deliverable that can inform downstream action.

4.1 Task Introduction

Given a policy-relevant question and a set of participant comments, the system produces a concise summary intended to serve as a deliverable: it should capture the range of perspectives and the reasons behind them, rather than collapsing the group into a single averaged voice.

4.2 Experiment Protocol

Experiment Data We experiment on DeliberationBank (Zhu et al., 2025), a benchmark containing U.S.-based public opinion comments spanning ten questions about technology, social media, and public policy.

Experiment Details We primarily evaluate teams of four participants. For each question, we cluster the crowd-sourced comments into four groups and sample one comment from each cluster, forming a team intended to cover qualitatively different stances. We sample a total of 500 team configurations. We follow DeliberationBank protocol (Zhu et al., 2025) to score the outputs, and in addition using LLM as a judge (Li et al., 2025) to perform pairwise comparison.

Baselines and Metrics We compare with: 1. **Direct summary** that prompts an LLM to summarize the comments, 2. **Chain-of-Thought** (CoT) that prompts LLM to think before generating a final summary (Wei et al., 2022), 3. **Self-Refinement** (Self-Refine), iteratively refining summaries without structured interaction (Madaan et al., 2023), 4. **Multi-Agent Debate** (MAD), using generic agents conducting four rounds of debate (Du et al., 2023). We report four dimensions from DeliberationBank: representativeness, informativeness, neutrality, and policy approval. A detailed description of the four metrics and their significance to open-ended decision making is presented in Table 7.

4.3 Experiment Results

TeamFusion consistently outperforms baselines. Table 1 shows consistent gains from TeamFusion across base models and question types. We focus on **representativeness** as the primary metric because it directly captures our goal of preserving diverse viewpoints. TeamFusion yields the largest improvements on representativeness, and these gains co-occur with strong increases in informativeness and policy approval, suggesting that the additional structured discussion surfaces missing considerations that make summaries more decision-ready. Importantly, neutrality remains comparable to baselines, indicating that improved viewpoint coverage does not come from introducing more polarized or editorial language.

To complement these aggregate scores, we also conduct a pairwise comparison between TeamFusion-generated summaries and direct summaries using an LLM-as-a-judge. We randomly sample 300 TeamFusion outcomes (100 for each base model), pair them with the corresponding direct summaries, and prompt GPT-4.1-mini to decide which summary is better. To avoid any po-

Model	Method	OpenQA				BinaryQA			
		Represent.	Inform.	Neutral.	Policy	Represent.	Inform.	Neutral.	Policy
Llama-3.3-70B	Direct	$.586_{.006}$	$.537_{.006}$	$.580_{.006}$	$.574_{.007}$	$.595_{.007}$	$.541_{.006}$	$.590_{.007}$	$.542_{.007}$
	CoT	$.568_{.007}$	$.524_{.006}$	$.570_{.006}$	$.561_{.007}$	$.578_{.007}$	$.530_{.006}$	$.580_{.006}$	$.528_{.007}$
	Self-Refine	$.577_{.007}$	$.536_{.006}$	$.566_{.006}$	$.570_{.007}$	$.599_{.007}$	$.545_{.006}$	$.587_{.007}$	$.547_{.007}$
	MAD	$.588_{.007}$	$.544_{.006}$	$.572_{.007}$	$.579_{.007}$	$.596_{.007}$	$.554_{.006}$	$.585_{.007}$	$.549_{.006}$
	TeamFusion	$\mathbf{.608}_{.007}$	$\mathbf{.588}_{.006}$	$\mathbf{.587}_{.006}$	$\mathbf{.602}_{.007}$	$\mathbf{.620}_{.007}$	$\mathbf{.597}_{.006}$	$\mathbf{.610}_{.007}$	$\mathbf{.568}_{.007}$
GPT-4.1-mini	Direct	$.582_{.006}$	$.534_{.006}$	$.576_{.006}$	$.579_{.007}$	$.589_{.007}$	$.531_{.006}$	$.584_{.007}$	$.543_{.006}$
	CoT	$.581_{.007}$	$.525_{.006}$	$.568_{.006}$	$.571_{.007}$	$.580_{.007}$	$.523_{.006}$	$.575_{.007}$	$.538_{.006}$
	Self-Refine	$.608_{.007}$	$.544_{.007}$	$.571_{.007}$	$.593_{.008}$	$.610_{.007}$	$.553_{.007}$	$.580_{.008}$	$.566_{.006}$
	MAD	$.598_{.007}$	$.553_{.006}$	$.569_{.006}$	$.592_{.008}$	$.616_{.006}$	$.559_{.007}$	$.586_{.008}$	$.570_{.006}$
	TeamFusion	$\mathbf{.614}_{.007}$	$\mathbf{.594}_{.006}$	$\mathbf{.585}_{.006}$	$\mathbf{.608}_{.007}$	$\mathbf{.623}_{.007}$	$\mathbf{.601}_{.006}$	$\mathbf{.604}_{.007}$	$\mathbf{.587}_{.006}$
GPT-4.1	Direct	$.578_{.006}$	$.531_{.006}$	$.573_{.006}$	$.575_{.007}$	$.584_{.007}$	$.530_{.006}$	$.577_{.006}$	$.541_{.006}$
	CoT	$.582_{.006}$	$.537_{.006}$	$.572_{.006}$	$.577_{.006}$	$.585_{.007}$	$.533_{.006}$	$.578_{.007}$	$.539_{.006}$
	Self-Refine	$.595_{.007}$	$.556_{.006}$	$.574_{.006}$	$.594_{.007}$	$.605_{.008}$	$.543_{.006}$	$.575_{.007}$	$.557_{.007}$
	MAD	$.599_{.007}$	$.561_{.007}$	$.576_{.007}$	$.594_{.007}$	$.609_{.010}$	$.563_{.006}$	$.580_{.008}$	$.567_{.006}$
	TeamFusion	$\mathbf{.621}_{.007}$	$\mathbf{.622}_{.007}$	$\mathbf{.582}_{.006}$	$\mathbf{.619}_{.007}$	$\mathbf{.640}_{.007}$	$\mathbf{.634}_{.006}$	$\mathbf{.602}_{.007}$	$\mathbf{.603}_{.007}$

Table 1: Performance comparison between TeamFusion and baselines on DeliberationBank task. We present the results on the two sub-categories of the questions: OpenQA and BinaryQA. Values are mean ± 95% CI. We report scores for representativeness (Represent.), informativeness (Inform.), neutrality (Neutral.), and policy approval (Policy; higher is better). Best scores per column are in **bold**.

	Win / Tie / Loss (%)			
Model	Represent.	Inform.	Neutral.	Policy
Llama-70B	71 / 28 / 1	95 / 0 / 5	51 / 43 / 6	97 / 0 / 3
GPT-4.1-mini	72 / 26 / 2	96 / 0 / 4	27 / 61 / 12	96 / 0 / 4
GPT-4.1	93 / 7 / 0	98 / 0 / 2	46 / 49 / 5	99 / 0 / 1

Table 2: Win/Tie/Loss rate of TeamFusion outcome against direct summary across four metrics. Higher win rates indicate stronger relative performance.

sitional bias of the LLM judge (Shi et al., 2024), we randomized the order of the summaries in the prompt. As shown in Table 2, TeamFusion wins overwhelmingly on informativeness and policy approval, and wins on representativeness in the large majority of cases.

Performance gains stem from personalized interaction. We compare TeamFusion against compute-matched Self-Refine and MAD baselines. While MAD involves structured discussion among agents, its generic approach without personalization leads to limited viewpoint diversity and narrower coverage. Similarly, self-refinement provides an iterative reasoning structure but lacks interactive discussion. In contrast, TeamFusion's performance boost is primarily attributable to its personalized agent interactions, highlighting the critical role of personalization and structured discussion. This analysis confirms that the observed improvements are not merely due to increased computational budget but rather due to the strategic combination of personalized representation and iterative debating.

Method	Represent.	Inform.	Neutral.	Policy
Team size: 6				
Direct	$.582_{.010}$	$.537_{.008}$	$.575_{.009}$	$.560_{.010}$
CoT	$.576_{.010}$	$.529_{.008}$	$.569_{.009}$	$.556_{.010}$
Self-Refine	$.607_{.010}$	$.561_{.009}$	$.576_{.009}$	$.581_{.010}$
MAD	$.600_{.009}$	$.573_{.009}$	$.579_{.009}$	$.587_{.010}$
TeamFusion	$\mathbf{.622}_{.010}$	$\mathbf{.617}_{.008}$	$\mathbf{.593}_{.010}$	$\mathbf{.608}_{.010}$
Team size: 8				
Direct	$.580_{.009}$	$.540_{.008}$	$.580_{.008}$	$.556_{.009}$
CoT	$.575_{.009}$	$.530_{.008}$	$.573_{.009}$	$.550_{.009}$
Self-Refine	$.603_{.009}$	$.568_{.008}$	$.578_{.009}$	$.576_{.010}$
MAD	$.605_{.009}$	$.570_{.008}$	$.580_{.009}$	$.578_{.010}$
TeamFusion	$\mathbf{.621}_{.009}$	$\mathbf{.627}_{.007}$	$\mathbf{.596}_{.009}$	$\mathbf{.609}_{.009}$
Team size: 10				
Direct	$.578_{.007}$	$.546_{.007}$	$.579_{.008}$	$.554_{.008}$
CoT	$.574_{.008}$	$.535_{.007}$	$.571_{.007}$	$.550_{.008}$
Self-Refine	$.604_{.009}$	$.574_{.008}$	$.578_{.009}$	$.576_{.010}$
MAD	$.602_{.009}$	$.572_{.007}$	$.580_{.008}$	$.579_{.009}$
TeamFusion	$\mathbf{.619}_{.008}$	$\mathbf{.637}_{.008}$	$\mathbf{.596}_{.009}$	$\mathbf{.608}_{.009}$

Table 3: Performance comparison between TeamFusion and baselines on the DeliberationBank task for different team sizes. Values are mean ± 95% CI. Best scores per column are in **bold**.

TeamFusion demonstrates gains across different team sizes. We then investigate the effectiveness of team size. We fix the total number of sampled team configurations to 100 and use GPT-4.1-mini as the backbone. Results are shown in Table 3. Across different team sizes, TeamFusion

	Represent.	Inform.	Neutral.	Policy
Model: Llama-3.3-70B				
Base	$.582_{.007}$	$.541_{.006}$	$.568_{.006}$	$.552_{.006}$
+ Iter 1	$.601_{.006}$	$.563_{.005}$	$.579_{.006}$	$.566_{.006}$
+ Iter 2	$\mathbf{.618}_{.006}$	$\mathbf{.581}_{.005}$	$\mathbf{.592}_{.006}$	$\mathbf{.581}_{.006}$
Model: GPT-4.1-mini				
Base	$.622_{.018}$	$.596_{.014}$	$.596_{.016}$	$.599_{.016}$
+ Iter 1	$.633_{.018}$	$.625_{.015}$	$\mathbf{.604}_{.014}$	$.618_{.018}$
+ Iter 2	$\mathbf{.637}_{.018}$	$\mathbf{.638}_{.015}$	$.599_{.015}$	$\mathbf{.619}_{.016}$
Model: GPT-4.1				
Base	$.630_{.018}$	$.615_{.014}$	$.591_{.016}$	$.606_{.016}$
+ Iter 1	$.646_{.019}$	$.655_{.014}$	$.599_{.015}$	$.635_{.017}$
+ Iter 2	$\mathbf{.655}_{.017}$	$\mathbf{.686}_{.014}$	$\mathbf{.603}_{.015}$	$\mathbf{.643}_{.018}$

Table 4: Performance of different models under iterative refinement. Values are mean ± 95% CI.

consistently outperforms baselines on the key representativeness metric. TeamFusion can improve representativeness by about 0.04 absolute gain over the baselines, with non-overlapping confidence intervals. This demonstrates the scalability and generalization capability of TeamFusion.

Iterative refinement brings gains. We fix the total number of sampled team configurations to 100 and run TeamFusion over up to three iterations (i.e., two refinements). The results of iterative refinement are shown in Table 4. For all models, adding iterative refinement brings gains to the representativeness and informativeness of the final summary across the iterations. Neutrality and policy alignment also improve, though with smaller margins, suggesting that additional rounds are particularly effective at surfacing missing considerations rather than merely smoothing tone.

5 Task 2: Visual Design

We then study TeamFusion in a *human-centric, multi-modal* workflow grounded in *real industry practice*.

5.1 Problem Motivation

Creative alignment is a significant pain point in professional design. Unlike close-ended tasks with objective "gold labels," design briefs are open-ended and subject to interpretation. This ambiguity introduces friction in industry practice: teams must expend significant effort negotiating trade-offs between aesthetics, brand tone, and constraints.

We motivate this task by empirically quantifying this friction. In our preliminary analysis of professional designers' preferences (detailed in Sec 5.4), we observed that experts given the exact same brief and assets exhibited remarkably low agreement on quality. In 70% of cases, agreement was indistinguishable from random chance. This validates that divergent interpretation is a natural and pervasive bottleneck.

5.2 Task Setup

Each scenario consists of a client brief clarifying the requirements for an advertisement design and a set of candidate ad thumbnails. A team of professional designers rank the candidates and provide justifications. TeamFusion runs proxy agent discussion and produces remixed design images intended to better align with designers' expressed constraints while making the underlying points of agreement and disagreement actionable for downstream selection. We evaluate whether TeamFusion generated images can replace original team's favorites.

5.3 Experiment Protocol

We introduce a human-in-the-loop protocol to evaluate TeamFusion. Unlike static benchmark evaluations, this protocol allows us to scale realistic team interactions while keeping professional designers as the ultimate ground truth for decision quality.

Phase 0: Realistic Scenario Construction We construct 50 high-quality design scenarios derived from real social media advertising campaigns (Yamaguchi, 2021). Each scenario includes a professional client brief and a set of diverse candidate designs. All scenarios have been validated by two external senior designers for realism. Full construction details can be found in Appendix G.2.

Phase 1: Preference collection We recruit 9 professional designers to annotate scenarios asynchronously. For each assigned scenario, designer ranks the six options and writes a brief justification. Each scenario received at least four independent annotations, serving as the "seed" evidence that conditions our proxy agents.

Phase 2: Simulation with nominal teams For each scenario, we form two nominal team settings from asynchronous annotations: **Full-Team** (all available annotations) and **Small-Team** (a random subset of two designers). We run TeamFusion for three iterations, producing one new remixed design per iteration.This yields 100 TeamFusion runs and 300 remixed design candidates. We also record

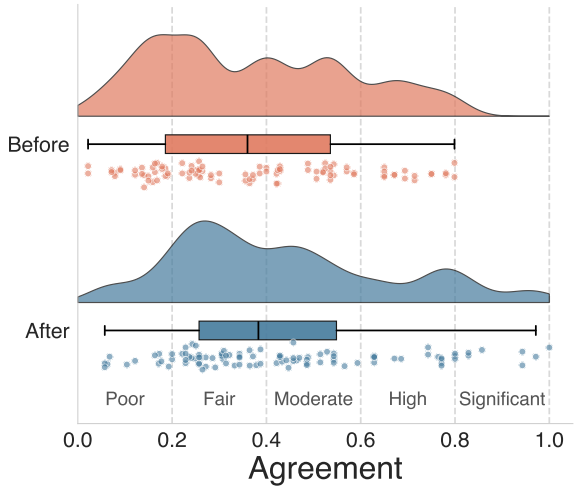

Figure 3: The distribution of agreement scores to measure dataset-wide agreement before and after TeamFusion's execution. The data is categorized into five value ranges to interpret agreement strength. Agreements across 100 team settings after running TeamFusion show a dataset-wide move towards higher agreement.

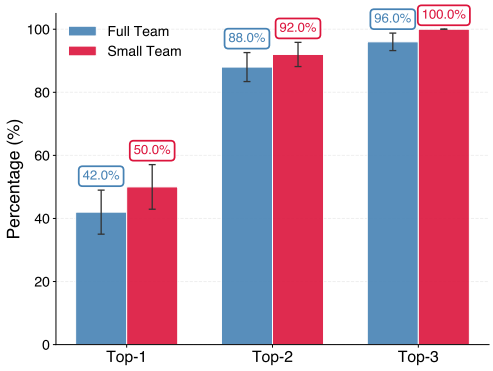

Figure 4: The rate of TeamFusion-generated images appearing in the final top-ranked selections. Error bars represent the 95% confidence interval.

each proxy agent's discussion comments about the generated candidates for later analysis.

Phase 3: Designer re-evaluation To determine if the system successfully facilitated convergence, we close the loop by returning the generated outputs to the original human designers. We combine the team's initial top three options via Borda count with the three TeamFusion-generated options. Designers then re-rank the combined set and rate whether their proxy agent's commentary aligns with their own reasoning.

5.4 Experiment Results

Our analysis reveals three main findings. First, we empirically verify the motivating problem of divergent preferences in teams. Second, we show that TeamFusion can generate consensus-oriented remixes that successfully induce convergence. Finally, we show that real designers largely agree with the debate commentary made by their delegate agents, indicating that simulated debates are well-grounded.

Finding 1: Divergent interpretations are a real, salient problem. As outlined in our motivation, we hypothesized that professional designers hold conflicting interpretations of the same brief. Our analysis of the pre-discussion ranking data confirms this friction is substantial. To quantify this, we calculate Kendall's Coefficient of Concordance (W) on the independent rankings provided in Phase 1. As shown in Figure 3 (top), the agreement among designers is consistently low, with a mean of 0.37 (falling into the "Fair Agreement" range). Notably, 84% of scenarios fall into the "Moderate" or lower agreement categories, and in 70% of cases, the agreement among professionals is not statistically significant ($p \geq 0.05$). This widespread lack of consensus in the real data confirms that our motivating problem is natural-arising and salient, providing strong empirical evidence in support of systems like TeamFusion.

Finding 2: TeamFusion can support team convergence by generating consensus-oriented designs. Our results reveal two ways in which TeamFusion-generated design revisions meaningfully modify the output of creative teams.

The results presented in Figure 4 show that **TeamFusion-generated options become the single top-ranked option across the team in nearly half of all test cases**, displacing the original team-wide favorites. Notably, this indicates that the execution of TeamFusion can be seen as a generative AI feature with significant team-wide acceptance rate under the strictest decision-making scenario, that is, the team decides to move forward with the single best design only.

Under less strict decision-making scenarios, TeamFusion also shows potential for contributing to teams' outputs. We find that TeamFusion-generated option appeared in the top-two rankings in a remarkable 88% of Full-Team and 92% of Small-Team test cases. This finding is noteworthy because, in the creative decision-making space, teams may use not only the single best option out of group ideation, but actually a short-list of top

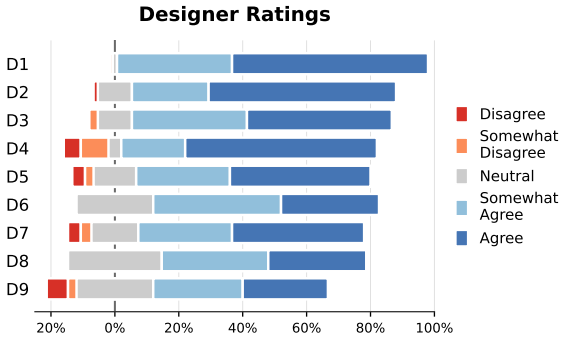

Figure 5: Distribution of annotator ratings for agreement with agent-generated commentary, grouped by designers. The results show an overwhelmingly positive perception.

options; for example, teams may steer the creative process by revising on one of them.

More globally, Figure 3 shows how there is a dataset-wide move towards higher agreement after the exposure to TeamFusion outputs: mean Kendall's W rises from 0.37 to 0.43. Qualitatively, the dataset mean moves from only "Fair Agreement" before to "Moderate Agreement" after TeamFusion, which further indicates the effectiveness of our system in supporting convergence.

Finding 3: Designers feel largely represented by their proxy agents. As shown in Figure 5, the results were **overwhelmingly positive**. The scores were strongly skewed toward agreement, with a mean of 4.06 ($\sigma = 1.07$). Over 75% of all comments received a positive score. This indicates that designers broadly perceived the outputs of their proxy agents as natural-sounding and representative of their own design rationales. This positive perception was highly consistent across our participants. An analysis of designer-level means showed a narrow range, with the lowest average rating being 3.53. This indicates that even the most critical participant found the commentary to be representative. The low standard deviation of these designer means ($\sigma = 0.29$) further reinforces that this high level of agreement is a shared, consistent finding.

5.5 Live User Study Results

To complement our asynchronous evaluation, we run a live, controlled within-team study to test whether TeamFusion facilitates convergence in an end-to-end workflow where participants create and revise designs. We recruit six participants and formed two teams of three, with each team completing two tasks in a counterbalanced crossover design[1]. Due to the small number of participants, we report this live study as a pilot with descriptive results rather than a statistically powered evaluation. As shown in Table 5, using TeamFusion leads to faster team decisions compared to free-form discussion, while also improving participants' perceived representativeness, clarity of trade-offs, and overall satisfaction with the team outcome. After experiencing both workflows, a strong majority of participants explicitly preferred TeamFusion over free-form discussion, showcasing the effectiveness over unconstrained collaboration.

Metric	Discussion	TeamFusion
Decision Time (min) ↓	18.0	**12.4**
Q1: Representative ↑	3.7	**4.3**
Q2: Clarity ↑	3.5	**3.8**
Q3: Satisfaction ↑	3.5	**4.2**
Preferred ↑	1/6	**5/6**

Table 5: Live within-team study results. Each team completed two briefs in a counterbalanced crossover design.

6 Case Study

We present a case study on the effectiveness in Figure 7. Due to page limit, we defer the detailed analysis in Appendix D. The core takeaway is that TeamFusion can better preserve fine-grained, participant-specific content than direct aggregation. We also present a case study on the failure mode of TeamFusion in Appendix D.2.

7 Conclusion

Open-ended team decisions require deliverables that make trade-offs and disagreements visible rather than averaging them away, yet common aggregation methods often erase minority or conditional viewpoints and reduce auditability. We addressed this challenge with TeamFusion, a multi-agent framework that shifts the paradigm from direct aggregation to modeled interaction. By representing participants with preference-grounded proxy agents and orchestrating structured debates, TeamFusion fully develops the consensus-seeking process to produce editable, rationale-backed deliverables. Evaluations on two teamwork tasks show that explicitly modeling interaction improves viewpoint coverage and decision usefulness and can

[1]Full details can be found in Appendix H

induce greater convergence, validating the potential of AI to facilitate human collaboration.

Limitations

Our findings show that TeamFusion can successfully support the creative convergence process, producing consensus-inducing design revisions grounded in natural-sounding, agreeable rationales. However, our work has limitations that shed light on important directions for future research. The current implementation assumes a flat hierarchy in the team, not accounting for different roles (e.g., art directors managing the team, or even clients themselves in the loop) and seniority levels (e.g., senior vs. junior designers). Even more realistic professional settings may warrant slightly different assumptions, with implications to our current modeling decisions.

Ethics Statement

TeamFusion is designed to **support**—and **not** substitute—creative teams in their decision-making, keeping human taste and judgment front and center in how the system is used and guided. The configuration of iterations and the production of iterative outputs further aligns with the principle that human teams are not only the users guiding the system but also the ultimate decision-makers determining whether and to what extent TeamFusion's outputs are useful. For the same reason, we propose that the underlying prompts producing the outputs and their provenance back to the multi-agent debates are kept transparent to users, such that they can be reviewed and remixed themselves, providing full human-in-the-loop control.

References

Selcuk Acar and Mark A Runco. 2019. Divergent thinking: New methods, recent research, and extended theory. *Psychology of aesthetics, creativity, and the arts*, 13(2):153.

Divyansh Agarwal, Alexander Fabbri, Ben Risher, Philippe Laban, Shafiq Joty, and Chien-Sheng Wu. 2024. Prompt leakage effect and mitigation strategies for multi-turn LLM applications. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1255–1275, Miami, Florida, US. Association for Computational Linguistics.

Adithya Bhaskar, Alex Fabbri, and Greg Durrett. 2023. Prompted opinion summarization with GPT-3.5. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9282–9300, Toronto, Canada. Association for Computational Linguistics.

Duncan Black. 1948. On the rationale of group decision-making. *Journal of political economy*, 56(1):23–34.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2006. The ami meeting corpus: A pre-announcement. In *Machine Learning for Multimodal Interaction*, pages 28–39, Berlin, Heidelberg. Springer Berlin Heidelberg.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Harrison Chase. 2022. Langchain. https://github.com/langchain-ai/langchain.

John Chen, Xi Lu, Yuzhou Du, Michael Rejtig, Ruth Bagley, Mike Horn, and Uri Wilensky. 2024. Learning agent-based modeling with llm companions: Experiences of novices and experts using chatgpt & netlogo chat. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*.

Steffi Chern, Zhen Fan, and Andy Liu. 2024. Combating adversarial attacks with multi-agent debate. *arXiv preprint arXiv:2401.05998*.

Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. 2024. Enhancing ai-assisted group decision making through llm-powered devil's advocate. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pages 103–119.

Luis Fariñas del Cerro, Andreas Herzig, Dominique Longin, and Omar Rifi. 1998. Belief reconstruction in cooperative dialogues. In *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*, pages 254–266. Springer.

Karen L. Dowling and Robert D. St. Louis. 2000. Asynchronous implementation of the nominal group technique: is it effective? *Decis. Support Syst.*, 29(3):229–248.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.

B Aubrey Fisher. 1970. Decision emergence: Phases in group decision-making. *Communications Monographs*, 37(1):53–66.

Jarod Govers, Eduardo Velloso, Vassilis Kostakos, and Jorge Goncalves. 2024. Ai-driven mediation strategies for audience depolarisation in online debates. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM workshop on artificial intelligence and security*, pages 79–90.

Xudong Guo, Kaixuan Huang, Jiale Liu, Wenhui Fan, Natalia Vélez, Qingyun Wu, Huazheng Wang, Thomas L Griffiths, and Mengdi Wang. 2024. Embodied llm agents learn to cooperate in organized teams. *arXiv preprint arXiv:2403.12482*.

Shanshan Han, Qifan Zhang, Weizhao Jin, and Zhaozhuo Xu. 2024. Llm multi-agent systems: Challenges and open problems. *arXiv preprint arXiv:2402.03578*.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, and 1 others. 2023. Metagpt: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*.

Zhe Hu, Hou Pong Chan, Jing Li, and Yu Yin. 2025. Debate-to-write: A persona-driven multi-agent framework for diverse argument generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4689–4703.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Nannan Huang, Lin Tian, Haytham Fayek, and Xiuzhen Zhang. 2023. Examining bias in opinion summarisation through the perspective of opinion diversity. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 149–161, Toronto, Canada. Association for Computational Linguistics.

Qiushi Huang, Xubo Liu, Tom Ko, Boyong Wu, Wenwu Wang, Yu Zhang, and Lilian Tang. 2024. Selective prompting tuning for personalized conversations with llms. *ArXiv*, abs/2406.18187.

Pontus Johansson. 2002. User modeling in dialog systems. *St. Anna Report SAR*, pages 02–2.

Sara Kiesler and Lee Sproull. 1992. Group decision making and communication technology. *Organizational behavior and human decision processes*, 52(1):96–123.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37:79410–79452.

Alfred Kobsa. 1989. A taxonomy of beliefs and goals for user models in dialog systems. In *User models in dialog systems*, pages 52–68. Springer.

Kenneth L Kraemer and John Leslie King. 1988. Computer-based systems for cooperative work and group decision making. *ACM Computing Surveys (CSUR)*, 20(2):115–146.

D. Kwon, Sunwoo Lee, Ki Hyun Kim, Seojin Lee, Taeyoon Kim, and Eric Davis. 2023. What, when, and how to ground: Designing user persona-aware conversational agents for engaging dialogue. In *Annual Meeting of the Association for Computational Linguistics*.

Philippe Laban, Wojciech Kryściński, Divyansh Agarwal, Alexander Richard Fabbri, Caiming Xiong, Shafiq Joty, and Chien-Sheng Wu. 2023. Summedits: Measuring llm ability at factual reasoning through the lens of summarization. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 9662–9676.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2025. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791.

Miao Li, Eduard Hovy, and Jey Lau. 2023. Summarizing multiple documents with conversational structure for meta-review generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7089–7112, Singapore. Association for Computational Linguistics.

Miao Li, Jey Han Lau, and Eduard Hovy. 2024. A sentiment consolidation framework for meta-review generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10158–10177, Bangkok, Thailand. Association for Computational Linguistics.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 17889–17904.

Xiaogeng Liu, Zhiyuan Yu, Yizhe Zhang, Ning Zhang, and Chaowei Xiao. 2024. Automatic and universal prompt injection attacks against large language models. *arXiv preprint arXiv:2403.04957*.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2025. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72.

R Orwig, Hsinchun Chen, D Vogel, and Jay F Nunamaker. 1997. A multi-agent view of strategic planning using group support systems and artificial intelligence. *Group Decision and Negotiation*, 6(1):37–59.

Letitia Parcalabescu and Anette Frank. 2024. On measuring faithfulness or self-consistency of natural language explanations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6048–6089.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. 2024. Cooperate or collapse: Emergence of sustainable cooperation in a society of llm agents. *Advances in Neural Information Processing Systems*, 37:111715–111759.

Steven Rogelberg, Desmond J. Leach, Peter B. Warr, and Jennifer L. Burnfield. 2006. "not another meeting!" are meeting time demands related to employee well-being? *The Journal of applied psychology*, 91 1:83–96.

Alexander C. Romney, Joseph A. Allen, and Zahra Heydarifard. 2025. Meeting load paradox: Balancing the benefits and burdens of work meetings. *Business Horizons*, 68(1):33–43.

Mark A Runco and Selcuk Acar. 2012. Divergent thinking as an indicator of creative potential. *Creativity research journal*, 24(1):66–75.

Donald G. Saari. 1995. *Basic Geometry of Voting*, 1 edition. Springer-Verlag Berlin Heidelberg.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, and 1 others. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2024. Judging the judges: A systematic study of position bias in llm-as-a-judge. *arXiv preprint arXiv:2406.07791*.

Joon Gi Shin, Janin Koch, Andrés Lucero, Peter Dalsgaard, and Wendy E. Mackay. 2023. Integrating ai in human-human collaborative ideation. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI EA '23, New York, NY, USA. Association for Computing Machinery.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 97–100, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Linxin Song, Jiale Liu, Jieyu Zhang, Shaokun Zhang, Ao Luo, Shijian Wang, Qingyun Wu, and Chi Wang. 2024. Adaptive in-conversation team building for language model agents. *arXiv preprint arXiv:2405.19425*.

Qiushi Sun, Zhangyue Yin, Xiang Li, Zhiyong Wu, Xipeng Qiu, and Lingpeng Kong. Corex: Pushing the boundaries of complex reasoning through multi-model collaboration. In *First Conference on Language Modeling*.

Michael Henry Tessler, Michiel A. Bakker, Daniel Jarrett, Hannah Sheahan, Martin J. Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C. Parkes, Matthew Botvinick, and Christopher Summerfield. 2024. Ai can help humans find common ground in democratic deliberation. *Science*, 386(6719):eadq2852.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D Nguyen. 2025. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, and 1 others. 2023. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521*.

Zheng Wang, Zhongyang Li, Zeren Jiang, Dandan Tu, and Wei Shi. 2024. Crafting personalized agents

through retrieval-augmented generation on editable memory graphs. In *Conference on Empirical Methods in Natural Language Processing*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2024. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*.

Kota Yamaguchi. 2021. Canvasvae: Learning to generate vector graphic documents. *ICCV*.

Miao Yu, Fanci Meng, Xinyun Zhou, Shilong Wang, Junyuan Mao, Linsey Pan, Tianlong Chen, Kun Wang, Xinfeng Li, Yongfeng Zhang, and 1 others. 2025. A survey on trustworthy llm agents: Threats and countermeasures. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 6216–6226.

Haopeng Zhang, Philip S Yu, and Jiawei Zhang. 2025. A systematic survey of text summarization: From statistical methods to large language models. *ACM Computing Surveys*, 57(11):1–41.

Shaokun Zhang, Jieyu Zhang, Jiale Liu, Linxin Song, Chi Wang, Ranjay Krishna, and Qingyun Wu. 2024a. Offline training of language model agents with functions as learnable weights. In *Forty-first International Conference on Machine Learning*.

Shaowei Zhang and Deyi Xiong. 2025. Debate4math: Multi-agent debate for fine-grained reasoning in math. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16810–16824.

Yang Zhang, Hanlei Jin, Dan Meng, Jun Wang, and Jinghua Tan. 2024b. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. *arXiv preprint arXiv:2403.02901*.

Yusen Zhang, Nan Zhang, Yixin Liu, Alexander Fabbri, Junru Liu, Ryo Kamoi, Xiaoxin Lu, Caiming Xiong, Jieyu Zhao, Dragomir Radev, Kathleen McKeown, and Rui Zhang. 2024c. Fair abstractive summarization of diverse perspectives. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3404–3426, Mexico City, Mexico. Association for Computational Linguistics.

Shenzhe Zhu, Shu Yang, Michiel A Bakker, Alex Pentland, and Jiaxin Pei. 2025. Can ai truly represent your voice in deliberations? a comprehensive study of large-scale opinion aggregation with llms. *arXiv preprint arXiv:2510.05154*.

A Discussion

In this section, we discuss the broader implications of our findings. We begin by examining the contributions of our user study procedure as a scalable method for evaluating AI systems for group ideation. We then consider the potential of TeamFusion's architecture as a generalizable model for AI-facilitated team convergence beyond graphic design. Finally, we address the limitations of our work and outline promising directions for future research.

A.1 User Study Procedure

One of the primary contributions of this work is the evaluation procedure itself. Research into AI systems for team settings, particularly in creative domains, is often hampered by methodological challenges and logistical overhead in human-in-the-loop evaluation. A key advantage of our three-phase protocol (Annotate $\rightarrow$ Simulate $\rightarrow$ Re-evaluate) is its scalability, which addresses this important bottleneck in team settings beyond text-only domains.

While existing work relies on live, synchronous sessions with participants, making them time-consuming, expensive, and difficult to scale beyond a small number of test cases, our approach is asynchronous and simulation-driven. By collecting designers' detailed rankings and justifications upfront (Phase 1), we effectively treat their expert judgment as a reusable resource. Instead of requiring designers to be online for every system execution, our approach can compose nominal teams from the offline annotations, simulate team dynamics, and return to team members to evaluate groundedness and effectiveness from system-generated deliverables. This allowed us to run 100 test cases covering a larger experimental space (i.e., two team sizes, three iterations of debating-and-remixing) much more efficiently. We hope the community interested in group ideation (Shin et al., 2023) can benefit from the key ideas behind our reproducible protocol.

A.2 AI as a Facilitator for Team Convergence in Other Creative Domains

While TeamFusion was implemented and evaluated within the domain of professional graphic design, its underlying architecture can be re-instantiated or extended to support team convergence in other creative domains. Advances in generative AI for video- or audio-editing, for example, pose interesting questions as to whether the positive findings we see in our studies would transfer to these other professional settings that similarly rely on group ideation.

A.3 Potential Risks

One significant risk involves the privacy and security implications of creating high-fidelity proxy agents conditioned on sensitive personal data. Since TeamFusion operates by encoding a participant's specific expressed preferences directly into a structured system prompt, there is an inherent risk that these digital proxies could inadvertently disclose more information than the user intended. For instance, while a user might strategically withhold certain views or "hidden assumptions" in a human-to-human setting, a proxy agent designed to "advocate for the assigned preference" might be manipulated via adversarial prompting (Greshake et al., 2023; Liu et al., 2024) or dialogue leaks (Agarwal et al., 2024) to reveal private rationales, biases, or competitive strategies to other agents in the shared "group chat" environment.

B Future Work

We are actively interested in exploring more hierarchical teams spanning different roles. Extending TeamFusion to other creative domains is another exciting direction that we identify, as well as further exploring "knobs" in the underlying parameter space such as the number of system iterations. Another important avenue of future work relates to more dynamic persona modeling: for example, agents could be designed to dynamically update their preferences and rationales based on the ongoing dialogue, better mimicking human adaptability and belief revision (Kobsa, 1989; Johansson, 2002; del Cerro et al., 1998).

A closely related challenge is resolving genuinely contradictory preferences within human teams. As will be discussed later in failure mode (Appendix D.2), when participants hold mutually exclusive positions, the remix phase can not resolve the conflict. This is not unique to our system; zero-sum disagreements are difficult even for human teams to settle without an explicit tie-breaking mechanism. We envision two complementary directions to address this. First, introducing hierarchical agent roles, where agents are assigned different levels of authority or domain expertise, could provide

a principled resolution path: for instance, when an Art Director and a Junior Designer disagree on a core brand element, the system could defer to the more senior role, mirroring how professional teams navigate such impasses in practice. Recent work on hierarchical multi-agent systems (Tran et al., 2025; Han et al., 2024; Guo et al., 2024) suggests that layered control structures can improve coordination without sacrificing the benefits of distributed deliberation. Second, rather than forcing the system to resolve every conflict autonomously, unresolved trade-offs could be surfaced back to the human team as interactive decision points (e.g., parameter sliders, toggle options, or side-by-side variant previews), letting the group make the final call on genuinely irreconcilable differences. This hybrid approach would preserve TeamFusion's ability to structure and externalize disagreement while keeping humans in control of decisions that require value judgments beyond what proxy agents can decide.

C Additional Results

C.1 Cost Analysis

We analyze the cost of running TeamFusion in this subsection.

For Task 1, we report the total cost for each task with GPT-4.1 as backbone. As shown in Table 6, TeamFusion is cheaper than both compute-matched baselines. MAD's generic agents receive the full summaries from all other agents, inflating token counts. Self-Refine regenerates the entire summary each iteration, compounding input length. In contrast, TeamFusion agents contribute to a shared group chat where message history builds incrementally, keeping per-turn context lean and cost-effective.

Method	Avg. Cost (10^{-2} USD)
Direct	0.12
CoT	0.19
Self-Refine	0.77
MAD	1.18
TeamFusion	**0.64**

Table 6: Cost comparison between TeamFusion and baselines on Task 1.

For Task 2, candidate generation costs about $0.90/scenario, and TeamFusion revision costs about $0.67/image, which is cost-effective given 42-50% top-1 acceptance. In terms of latency, our live study shows that TeamFusion reduces decision time from 18.0 to 12.4 minutes compared to free-form discussion, while also yielding higher representativeness and satisfaction scores.

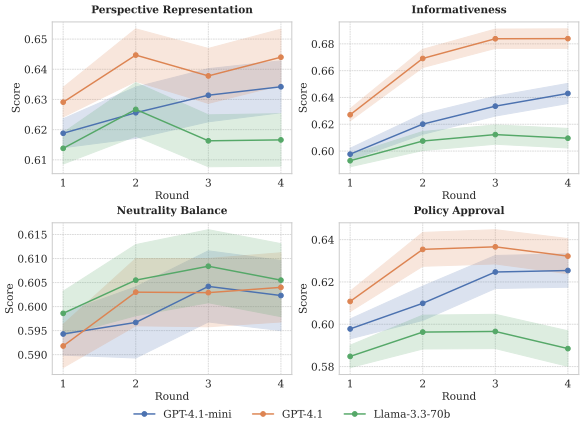

Figure 6: Ablations of per agent speaking turns on the four metrics.

C.2 Ablations on Per Agent Turns

We ablate on the number of discussion rounds in Task 1. We set the number of discussion rounds and observe TeamFusion performance across the four metrics. As shown in Figure 6, scaling up the rounds of discussion can improve informativeness and neutrality. Stronger proprietary models like GPT-4.1-mini and GPT-4.1 benefits more from increasing discussion rounds, representativeness increases. This showcases that adding more discussion can make different voices heard for stronger models. For weaker model Llama, adding round from 3 to 4 yields a decrease in four metrics. We hypothesize that this is due to its worse long context processing capability.

D Case Study

D.1 Case Study on Successful Outcomes

Figure 7 illustrates how TeamFusion better preserves fine-grained, participant-specific content than direct aggregation on a civic synthesis example. While the direct summary is fluent and captures the dominant positive sentiment (e.g., creativity, personalization, and efficiency), it noticeably homogenizes key details: it fails to carry forward Comment #1's concrete comparative claim that AI has "replaced" traditional search engines, and it collapses Comment #3's domain-specific experience ("my healthcare setting") into generic "work-related problems," obscuring where and why the

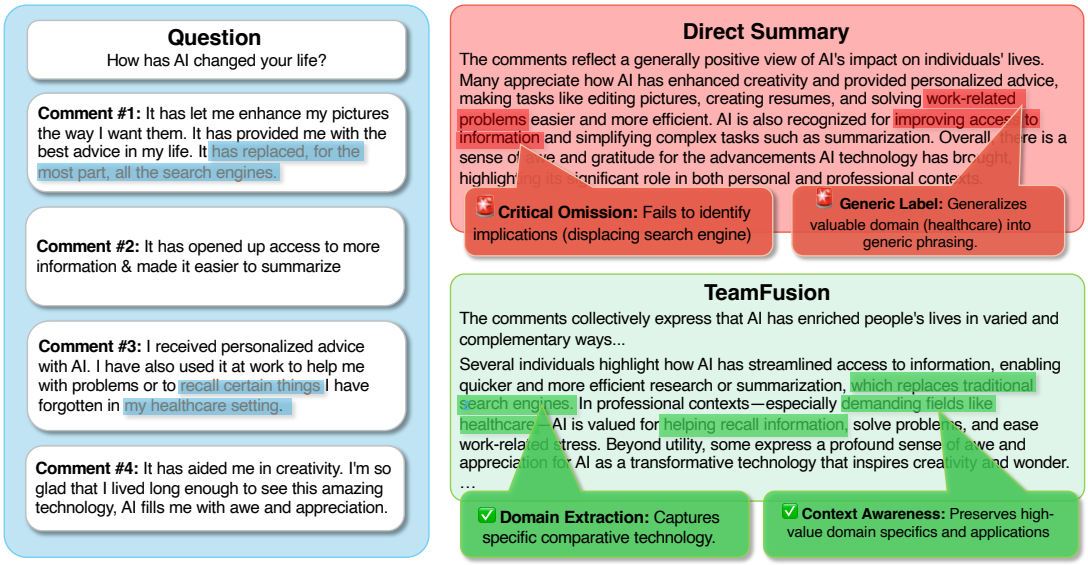

Figure 7: Case study in Task 1 comparing TeamFusion and direct summary. We partially omit outputs from TeamFusion to increase presentation focus.

tool is valuable. In contrast, TeamFusion output retains these high-salience details in the final deliverable. This example highlights two practical advantages of TeamFusion for generating deliverables: (i) minority or specialized experiences remain visible rather than averaged away, and (ii) concrete comparative statements and applications are maintained to support downstream interpretation and action. Overall, the case study qualitatively supports our quantitative gains on representativeness by showing that TeamFusion more faithfully carries forward what each participant uniquely contributed, instead of compressing distinct voices into a generic narrative.

D.2 Case Study on Failure Mode

While TeamFusion demonstrates strong performance across our evaluations, we identify a recurring failure pattern that reveals a fundamental limitation of the framework. From our qualitative inspection, TeamFusion underperforms when participants hold truly contradictory preferences about a specific topic. For example, in the visual design scenario, Designer A like a particular visual element while Designer B explicitly opposes it. Unlike cases where preferences differ in degree or emphasis, such zero-sum conflicts present an irreconcilable tension: any deliverable that satisfies one participant necessarily violates the other's stated constraint.

In these cases, the remix phase is forced to make a selection between the two conflicting viewpoints, leading to lower perceived representativeness for the team. This pattern is consistent with findings in the fair summarization literature, where aggregation over opposing stances risks producing outputs that no individual stakeholder endorses (Zhang et al., 2024c).

It is worth noting that this limitation is not unique to TeamFusion. Zero-sum disagreements are inherently difficult even for human teams to resolve without an explicit tie-breaking mechanism such as authority, voting, or external arbitration (Fisher, 1970; Black, 1948).

E Implementation Details of Task 1

E.1 Details of Represent

We design a prompt that consists of goal of the discussion, conversation style constraints, and preference samples. Prompt content is available at Appendix K.1.1.

E.2 Details of Remix

The remixing agent is a text-based LLM that takes in the task context, original comments, the discussion transcript, and generates a structured summary. Prompt content can be found at Appendix K.1.2.

E.3 Details of Iterative Revision

After an initial summary has been generated, the summary becomes a shared group message content as part of the task context. The agent group's collective goal changes to improve the summary to better reflect their individual standpoint. They

engage in the structured discussion to achieve the goal. Finally, the remixing agent ingests the individual comments based on the previous summary and the debate transcript, and generates a refined summary of the comments. The newly generated summary then becomes the summary to improve upon in the next round, if any.

E.4 Hyperparameters

We use three LLMs as backbones for both proxy agents and the remixing agent, i.e. Llama-3.3-70b, GPT-4.1-mini, GPT-4.1. We set the decoding temperature to 1. We set the number of per-debate turns to 1 in the main experiments. The agents are implemented and orchestrated by the AutoGen framework (Wu et al., 2024).

F Implementation Details of Task 2

F.1 Details of Representation

Based on recent works on LLM-based personalization (Chen et al., 2024; Huang et al., 2024; Wang et al., 2024; Kwon et al., 2023), we leverage in-context learning to customize an LLM agent for the participant. We combine best practices from prior work to compose a layered prompt that makes an LLM adhere to a given designer's preferences and reliably act as individual Designer Agents in our multi-agent system:

- **Overarching goal:** The first component of the prompt defines the overarching goal and the role of the agent. It establishes the agent's role as a 'design expert' and its goal is to reach consensus over potentially varying preferences.
- **Domain constraints:** Following the overarching goal, this component helps to concretize the "best practices" in communication that make the agent specifically natural-sounding and useful for design ideation. By prompting along more specific dimensions such as tone, language, and conciseness (e.g., *"Mimic real designers' tone and language style... Avoid very long messages"*), we further align the agent's output space to a desired communication style. This is expected to make agents' outputs more natural-sounding when verified by human evaluators.
- **Role-playing definition:** This component is the core of the personalization, explicitly limiting the agent's behavior to role-playing the human designer. The instruction, *"You are role-playing {user_name}. Always respond from the following perspective and expertise,"* acts as a powerful anchor that instructs the model to forego its default, neutral stance and instead adopt the specific viewpoint, expertise, and potential biases of the individual it represents, becoming their debate proxy.
- **Few-shot preference examples:** Finally, to concretely ground the agent on a preference set, the prompt is completed with the designer's opinions over the option space. These opinions include both ordered preferences—each option's ranking from best to worst according to that designer—as well as brief *ranking justifications* in natural language (e.g., *Image 4: {'rank': '1', 'justification': "It's bold and colorful, but feels more like a fashion brand than perfume. The bottle's squeezed in and doesn't really pop."}*). Considering the multi-modal nature of the input, which includes images and text, this component provides rich information for preference-grounding: the rank placements are explicit; the natural language justifications articulate explicit rationales; and even second-order preferences can be inferred from the image-text pair.

F.2 Details of Discussion

In Task 2, the overarching goal of the agent discussion is two-fold: (1) Reaching a consensus on the rankings of the six images according to adherence to the client brief and aesthetics; (2) Converging on the direction to improve based on the existing images.

F.3 Details of Remixing Agent

Once all proxy designer agents have reached the number of turns per debate, TeamFusion moves to translating the discussion outcomes into a revised image design. The remixing agent first consumes the entire chat history and extract two structured outputs: (1) The top-ranked options capturing the group's consensus, and (2) A set of remixing instructions, specifying which strengths from the top-ranked options to keep while addressing their weaknesses. The remixing agent then feeds the top-ranked options and the set of remixing instructions into a downstream image editing model. Importantly, this is a fundamental difference between

a group brainstorming technique such as nominal group technique (Dowling and St. Louis, 2000), which would apply voting at the end to obtain the top-ranked options, and TeamFusion's integration of generative AI in a consensus-oriented manner, incorporating AI as a creative partner.

F.4 Details of Iterative Revision

If the team would like to run another iteration on TeamFusion, the system is designed to narrow the scope on the top-ranked options and iteratively refine them. An initial option space with six designs is narrowed down to the three top-ranked options after the first debating iteration. A new remixed option is then added to this top three during remixing, finishing the first iteration with a top four. A second iteration would start from this top four, with the Designer Agents debating them for the same number of per-debate turns, narrowing down to a top two. A new remixed option would yield a top three at the end of the second iteration. Once iterative refinement is over, a final discussion yields the single best option for that run of TeamFusion.

F.5 Hyperparameters

We use GPT-4o as the backbone for the agents, with decoding temperature set to 1. The agents are implemented and orchestrated by the AutoGen framework (Wu et al., 2024). We leverage GPT-Image-1 for image remixing part of the Remixing Agent. Empirically, we have found that setting the number of per-debate turns to 2 allows agents in TeamFusion to negotiate in-depth without repeating themselves sycophantically (Sharma et al.), with little practical utility, or going off-topic—so we have fixed this parameter to 2 in this task. We run TeamFusion with 2 follow-up iterative revision rounds.

G User Study Details

G.1 Participants

We recruited 9 professional graphic designers (6 female, 3 male) on the Upwork platform. Each designer had substantial experience in social media ad design as recorded on the platform (mean jobs completed = 76.56, $\sigma = 95.18$), having successfully completed at the very least 10 projects. The compensation for participation varied by designer (mean = \$317.22, $\sigma = \$108.69$). Participants were anonymized and all procedures were approved by institutional review board.

G.2 Scenario Construction

We constructed the scenario in a three-phase process. First, we sampled 70 social media ads from the Crello dataset (Yamaguchi, 2021), filtering by the "Facebook Ad" and "Instagram Ad" categories. Second, we employed GPT-4o to reverse-engineer a hypothetical client brief for each ad image. Third, these client briefs and Crello ad images were sent into an image generation pipeline to create five new design variants for each setting. Specifically, the pipeline begins with a LLM planner that reads the image and the variation requirement, generates a plan to change certain aspects of the image. The plan is fed to a prompt writer LLM that consolidates the plan into a detailed instruction. GPT-Image-1 reads the instruction and the original image, then generates the image variant.

To validate the professional quality of client briefs and design options, we recruited two senior designers on Upwork who were native English speakers and had particularly extensive track records in dealing with real clients (311 and 520 completed projects). The designers scored all client briefs and design options on a 5-point Likert scale (1 = Completely Unrealistic, 5 = Fully Realistic)[2], allowing us to select 50 high-quality social media ad scenarios with positive scores from both judges, each including a realistic client brief and six design options.

G.3 Procedure

The relative scarcity of research addressing AI systems in team settings, particularly in the graphic design domain, motivated us to plan a novel procedure for composing teams of professional graphic designers. At a high level, this procedure consists of assigning different designers to the same setting (i.e., a client brief + six options), collecting their initial preferences over the option space, composing nominal teams to compute team-wide preferences, simulating these teams on TeamFusion, including TeamFusion-generated designs when collecting their new individual preferences, and returning to the nominal teams to compute new team-wide preferences, thus evaluating TeamFusion's contributions to the teams' final preferences. In detail, this procedure was implemented in three phases:

Phase 1: Initial designer annotation. Design-

[2]Full instructions can be found in the supplemental materials.

ers reviewed each brief and its six associated design options.[3] They were asked to rank the options based on how well they—subjectively—felt that each option addressed the client brief. For each ranking decision, they provided short (i.e., 2-3 sentences) written justifications expressing their judging rationales. We ensured that each social media ad scenario (client brief + six options) received at least four independent annotations, with each designer participating in exactly 25 scenarios.

Phase 2: TeamFusion run. For each of the 50 scenarios, we experimented with two team settings: (a) *Full-Team*, where all available designer data for a scenario were used to initialize Designer Agents; and (b) *Small-Team*, where random subsets of two designers were used to initialize Designer Agents. TeamFusion was fully executed for both the Full- and Small-Team settings, through three iterations of debating-and-remixing, collecting the TeamFusion-generated design at each iteration. As a result, we executed TeamFusion $50 \times 2 = 100$ times, and collected a total of $100 \times 3 = 300$ TeamFusion-generated options. From each execution, we also collect the simulated comments (per Section 3.4) that each Designer Agent makes for the three TeamFusion-generated options—all unseen by the original annotators.

Phase 3: Designer re-evaluation. For each of the 100 team settings, we collected the team's top three options using Borda count (Saari, 1995) from their initial rankings. We then mixed the three TeamFusion-generated options with the initial top three, for a new set of six options. By ranking TeamFusion's outputs vs. the initial top-ranked options, we can measure if—and to what extent—TeamFusion is able to modify the teams' top-ranked preferences. After being exposed to the unseen options in the re-ranking task, designers then score on a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree) their agreement with the simulated comments, measuring if—and to what extent—they feel represented by their proxy agents.

G.4 Participant Task Instructions

We present the task briefing to the participants in Figure 8, 9 and 10.

[3] Full instructions can be found in the supplemental materials.

Design Evaluation for Social Media Ad Annotation (Test Project)

Overview

- For the Design Evaluation job, design experts are asked to provide a small portfolio representing their work.
  - Specifically, this means providing 5 previously designed Social Media Ads paired with a short paragraph (50-100 words) detailing the client's original brief. It is totally fine to use personal projects where the designer imagined a hypothetical client.
- For each of a set of 25 Social Media Ads (either for Facebook or Instagram), designers are provided a client brief (around 150 words) and 6 diverse design options and are then asked to evaluate how well each option satisfies the client brief.
  - Specifically, this means ranking the 6 design options in order of preference, while also providing short comments (2-3 sentences) justifying each ranked position.

Detailed Scope

- Designers will be provided an Excel spreadsheet with two tabs: one named `task1_small_portfolio`, with 5 rows; and another named `task2_rank_and_justify`, with 25 rows.
- For the 5 rows in `task1_small_portfolio`, designers must populate two columns:
  - One named `Client Brief`, to be populated with a short paragraph (50-100 words) on that design's original brief.
  - And another named `Approved Design`, with the final approved design.
  - Again, it is totally fine to use personal projects where the designer imagined a hypothetical client.
- For the 25 rows in `task2_rank_and_justify`, designers will find 7 pre-populated columns, with both the client brief (around 150 words) and 6 diverse design options; as well as 7 columns to be populated by the designer.
  - To illustrate with an example:
    - Column `Client Brief`, such as:

Creative Brief for Instagram Ad

Project Overview:
Fleur de Sel, a luxury fragrance brand, aims to promote its new perfume through an Instagram ad. With a reputation for elegance and refinement, Fleur de Sel competes with premium brands like Chanel and Dior. Our goal is to capture the essence of sophistication and allure, showcasing the perfume as a must-have for the modern woman.

Objectives:
Launch the campaign by next month with a target of reaching 100,000 impressions and increasing web traffic by 20%. The ad should foster brand awareness and drive conversions through social media engagement.

Target Audience:
Women aged 25-40, primarily middle to upper class, who value luxury and style. They are trend-conscious and frequent users of high-end fashion and beauty magazines. Their needs include finding a signature scent that embodies their elegant lifestyle.

Messaging:
Communicate the allure of the new fragrance with the message: "Feel the Divine Breeze of the Sea". Employ a sophisticated and elegant tone to evoke a sense of exclusiveness and refinement.

Deliverables:
Create an Instagram ad in square format (1080x1080 pixels), delivered in JPG and PSD formats. Ensure visual aesthetics are on-brand with a focus on black and white imagery to emphasize elegance.

    - Column `Design Option #1`, such as:

    - Column `Design Option #2`, such as:

    - Column `Design Option #3`, such as:

  - Column `Design Option #4`, such as:

Figure 8: Task instruction part 1.

Figure 9: Task instruction part 2.

- Column `Design Option #5`, such as:

- Column `Design Option #6`, such as:

- The designer must populate the following columns:
  - Column `Rank` must indicate how each numbered design option on the sheet ranks from best to worst in satisfying the brief, according to the designer's preferences (for example: 1,2,6,5,4,3).
  - And `Justification 1`, `Justification 2`, `Justification 3`, `Justification 4`, `Justification 5`, and `Justification 6` must be populated with brief comments (2–3 sentences) justifying each ranked position (for example: This ad has nice quality photography but the dress the model is wearing seems out of place for the subject. The first you notice is the dress, not the perfume. And it doesn't communicate sophistication.).

Figure 10: Task instruction part 3.

H Live User Study Details

H.1 Goal

We conducted a small live study to evaluate TeamFusion in an end-to-end collaborative design workflow where team members (i) create initial candidate ad thumbnails using generative tools, (ii) express individual preferences via rankings and rationales, and (iii) collaboratively converge on a final selection through either TeamFusion or through free-form discussion and revision. The study focuses on decision-process outcomes and participant-reported experience.

H.2 Experiment Protocol

We recruited 6 participants and formed 2 teams of 3. Each team completed two ad-brief tasks (Brief A and Brief B). We **counterbalanced condition order** at the team level: (i) Team 1 used TeamFusion on Brief A and the baseline on Brief B; (ii) Team 2 used the baseline on Brief A and TeamFusion on Brief B. This controls for brief-specific difficulty and order effects.

H.3 Materials

Ad briefs. We prepared two ad briefs following the same protocol as G.2.

Reference gallery. For each brief, we provided a gallery of 10 image thumbnails as references for generation. This is to seed stylistic directions and reduce cold-start variance in what participants create.

Generative tools. All participants had access to the same GenAI editing and generation tools, specifically GPT image editing and Nano-Banana.

H.4 Conditions

TeamFusion After participants provided their initial rankings and short rationales, we constructed one proxy agent per participant using these preference signals. TeamFusion then produced: (1) a structured summary of agreements, disagreements, and key trade-offs; (2) two revised candidate thumbnails. Participants could optionally decide up to 3 additional refinement rounds by providing brief feedback (e.g., "strengthen product visibility").

Free-form discussion Participants coordinated through Zoom to discuss freely. They posted revisions through a private channel on Discord. They could use the tools to propose revised thumbnails and post them to the channel.

H.5 Procedure

Each brief followed the same phases:

Phase 0: Training (10 minutes) Participants first get aquainted with each other, then completed a short tutorial on the interface and tools by looking at admin-provided video demo.

Phase 1: Individual creation (15 minutes) Using the reference gallery and GenAI tools, each participant produced two initial candidate thumbnails. Participants uploaded their candidates to a shared board.

Phase 2: Individual preference elicitation (10 minutes). Participants independently ranked all initial candidates from best to worst for the brief and provided short justifications describing their key criteria.

Phase 3: Collaborative revision and convergence (20 minutes). Participants then completed one of the two conditions:

- **TeamFusion:** We ran TeamFusion using the Phase 2 evidence, producing revised designs and a structured trade-off summary. Participants reviewed the output and either (i) stopped and moved to Phase 4, or (ii) provided

brief feedback to trigger at most 3 additional refinement rounds.

- **Free-form discussion and revision:** Participants read each other's preference, and discussed freely in the voice channel and posted revisions they generated. The team stopped when time elapsed or when they agreed on a final candidate set.

Phase 4: Post-task questionnaire (1 minute). After the study, participants answered three 1–5 Likert items (1=strongly disagree, 5=strongly agree):

- **Q1 (Representativeness):** The final output reflected my key preferences and reasoning.
- **Q2 (Clarity):** It was clear what the main agreements, disagreements and trade-offs were and why.
- **Q3 (Outcome satisfaction):** I am satisfied with the final design decision our team reached.

After completing both briefs, participants answered a preference question: *"Which workflow would you choose for similar tasks?"* (TeamFusion/ Free-form discussion).

H.6 Study Interface

We present the screen shots of the UI used in live user study in Figure 11, 12. The UI enables users to upload their designs, writing critiques, reading each other's preferences, and monitoring TeamFusion output. The screenshot uses mock data for visualization purposes.

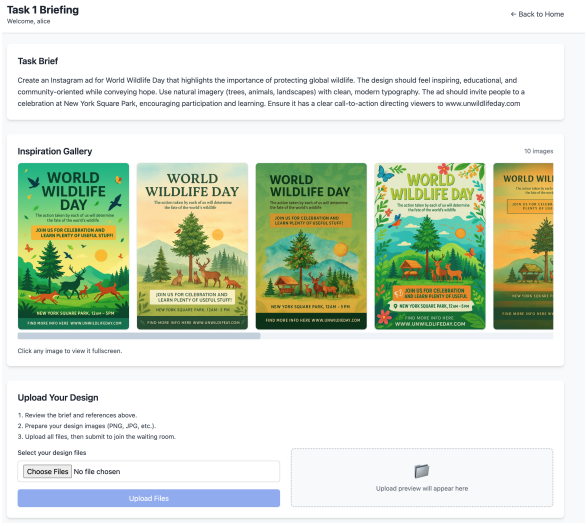

Figure 11: Screenshot of the live user study interface. This is the user upload design page.

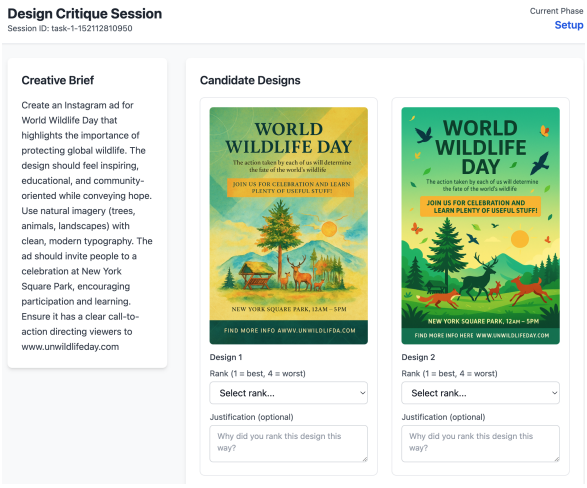

Figure 12: Screenshot of the live user study interface. This is the user critique design page.

I Detailed Explanation of Metrics

In Table 7, we describe the four metrics and explain why they matter for evaluating open-ended team decisions.

J AI Usage Disclosure

AI assistants were used to help polish the manuscript's wording and readability and to assist with drafting code snippets. All AI-assisted outputs were reviewed, verified, and edited by the authors.

K Prompts

In this section, we list all the prompts used in the experiments.

Metric	What it measures in this task	Why it matters for open-ended team decisions
Representativeness	Whether the summary covers the range of participant viewpoints and attributes key reasons/claims to the underlying comments (i.e., avoids viewpoint erasure).	Open-ended decisions require a deliverable that participants recognize as "their" perspectives being present. High representativeness reduces minority suppression and makes disagreement auditable rather than implicitly averaged away.
Informativeness	Whether the summary preserves concrete, decision-relevant content (rationales, constraints, trade-offs, edge cases), instead of generic paraphrases.	Teams need a deliverable that supports action: identifying what information would change a decision, what trade-offs are being made, and what constraints are binding. Higher informativeness makes the deliverable usable for follow-up discussion and planning.
Neutrality	Whether the summary maintains a balanced, non-editorial tone and avoids injecting the summarizer's own stance.	In civic/team settings, the deliverable often serves as shared ground for discussion. Neutrality helps prevent the system from "deciding" for the group via framing effects, preserving legitimacy and trust in the deliverable.
Policy approval	Whether the deliverable supports downstream acceptability/action (e.g., framing options in a way that is coherent, feasible, and aligned with the decision question).	Open-ended deliverables are judged not only by coverage, but by whether they help a team move forward. Policy approval captures whether the synthesized output is decision-oriented rather than merely descriptive.

Table 7: Evaluation dimensions for civic comment synthesis and their decision-support interpretation. We follow DeliberationBank's four metrics and interpret them as complementary requirements for open-ended team deliverables.

K.1 Task 1 Prompt

K.1.1 Prompt for Proxy Agent

```
Your name is {name}. You are a participant in a public deliberation discussion. Your role is to advocate for and discuss the perspective expressed in your assigned comment.
  ## Discussion Context **Question:** {question}
  **Your Assigned Comment:** {comment}
  ## Your Role and Instructions
  1. **Understand and Hold Your Position**: Carefully read and internalize the viewpoint expressed in your comment. This represents your perspective in this discussion. Stay true to the sentiment and reasoning of your assigned comment.
  2. **Advocate Effectively**:
  - Express the key points and reasoning behind your position
  - Always speak in concise and at most 2 paragraphs. Go straight to the core point.
  - Avoid adding any additional personal information or experience into discussion aside from given comments.
  3. **Engage Constructively**:
  - Listen to and acknowledge other participants' viewpoints
  - Identify common ground where it exists
  - Respectfully challenge points you disagree with, using reasoning and evidence
  4. **Contribute to Comprehensive Understanding**: Help ensure that your perspective is clearly understood and represented in the broader discussion, especially if it represents a minority or less common viewpoint.
  Remember: The goal is not to "win" the debate, but to ensure all perspectives—including minority opinions—are thoroughly heard, understood, and considered in the final summary of the deliberation.
```

K.1.2 Prompt for Remix Agent

```
You are summarizing a collection of comments for a deliberation question: {question}. You will first receive the comments. Then, a discussion between people who wrote the comments will follow. You must focus on comprehensively summarizing the comments and use the discussion to better understand the viewpoints of the comments. Please do not mention the total number of comments. Do not refer to any specific comment in the summary. If you need to provide statistical information, use percentages instead of absolute numbers.
  Here are the comments:
  {comments_str}
  Here is the discussion, use it to better understand the comments:
  {history_str}
```

K.2 Task 2 Prompt

K.2.1 System Prompt for Proxy Agent

The system prompt for Designer Agent has been presented and discussed in Section F.1.

```
You are a design expert participating in a discussion about design images. You have been given specific preferences over the designs and will discuss them with other experts to reach consensus. Advocate for your preferred designs while being open to other perspectives. Mimic how real designers' tone and language style to write to each other. Be concise and to the point. You are roleplaying as {user_name}. Always respond from the following perspective and expertise. Attached are the images paired with the justification, roleplay as if this is your preference. {formatted_preference}
```

K.2.2 Prompt for Discussion

```
Welcome to the design discussion! Each of you has seen the same set of images but may have different
```

preferences. Please discuss efficiently and work toward consensus on:

1. RANKING: Establish a ranked list of images from best to worst, considering both aesthetic appeal and alignment with the creative brief.

2. DESIGN IMPROVEMENT: Discuss how to enhance and combine the best elements from top 3 performing images. Consider:

- Primary composition and layout structure from the strongest images

- Visual elements that should be integrated or refined

- Color schemes and typography that work best

- Specific adjustments needed to balance different concerns

3. SYNTHESIS: Develop a cohesive approach that merges strengths from the top 3 performing images while addressing any weaknesses identified in the discussion. When you propose changes to improve, ground the instructions on top 3 performing images.

Share your reasoning and be open to different perspectives as you work toward both a final ranking and concrete design improvement directions.

Here is the creative brief for the task:

{brief}

K.2.3 Prompt for Remixing Agent

You are a design summarization expert analyzing a roundtable discussion between design experts about image variants. Your task is to carefully read through the entire conversation and extract two key outputs:

1. **FINAL RANKING**: Identify the consensus ranking of images from best to worst

2. **EDITING DIRECTIONS**: Extract specific instructions for creating an improved design by combining elements from different images

Analysis Instructions:

- Start from the END of the conversation and work backwards - the most recent messages contain the final consensus and should be given the highest priority. Early messages may contain initial disagreements or positions that were later changed.

- Focus on extracting the ultimate agreements on rankings and specific design recommendations that emerged at the conclusion of the discussion.

Requirements for editing_directions string:

Write detailed instructions as if directing an AI image editing model. It should include the following fields, if mentioned. If the discussion does not touch on the relevant field, don't include the field in your instruction.

1. **Primary Composition**. Example templates include:

Use the overall layout and structure from Image [number], specifically [describe the compositional elements, positioning, or arrangement].

2. **Visual Elements Integration**. Example templates include:

- Incorporate [specific visual element] from Image [number], such as [detailed description]

- Add [specific design feature] from Image [number], particularly [detailed description]

- Include [specific element] from Image [number], focusing on [detailed description]

3. **Color and Typography Refinements**. Example templates include:

- Adopt the [color scheme/typography style] from Image [number], specifically [details]

- Modify [specific aspect] using the approach seen in Image [number]

4. **Final Adjustments**. Example templates include:

- Ensure [specific requirement based on discussion]

- Balance [specific concern raised in discussion]

- Maintain [specific positive aspect mentioned]

Important Guidelines:

- Always reference images by their specific numbers (Image 1, Image 2, etc.)

- Be concrete and specific about visual elements (colors, positioning, typography, objects, etc.)

- Avoid vague language - use precise descriptions

- Focus on actionable instructions that an image editing AI could follow

- Only include elements and instructions that were actually discussed and agreed upon, never add your own novel thoughts

- If no clear consensus was reached, state this explicitly

Example 'editing_directions': "Incorporate the bold red CTA button from Image 3, positioning it in the lower-right corner as seen in Image 1, while maintaining the clean white background and centered product placement from Image 5."

Remember: Your output should be directly usable by downstream image editing systems, so precision and specificity are crucial.

Output Format:

You must output your analysis in the following JSON structure:

```json
{
  "final_ranking": [
    {
      "rank": 1,
      "image_number": <image_number>,
      "reason": "<reason_for_ranking>"
    },
    {
      "rank": 2,
      "image_number": <image_number>,
      "reason": "<reason_for_ranking>"
    },
    {
      "rank": 3,
      "image_number": <image_number>,
      "reason": "<reason_for_ranking>"
    }
  ],
  "editing_directions": "<instructions>"
}
```