DIALOGUE AS DISCOVERY: NAVIGATING HUMAN INTENT THROUGH PRINCIPLED INQUIRY

Anonymous authors
Paper under double-blind review

ABSTRACT

A fundamental bottleneck in human-AI collaboration is the "intention expression gap," the difficulty for humans to effectively convey complex, high-dimensional thoughts to AI. This challenge often traps users in inefficient trial-and-error loops and is exacerbated by the diverse expertise levels of users. We reframe this problem from passive instruction following to a Socratic collaboration paradigm, proposing an agent that actively probes for information to resolve its uncertainty about user intent. we name the proposed agent Nous, trained to acquire proficiency in this inquiry policy. The core mechanism of Nous is a training framework grounded in the first principles of information theory. Within this framework, we define the information gain from dialogue as an intrinsic reward signal, which is fundamentally equivalent to the reduction of Shannon entropy over a structured task space. This reward design enables us to avoid reliance on costly human preference annotations or external reward models. To validate our framework, we develop an automated simulation pipeline to generate a large-scale, preference-based dataset for the challenging task of scientific diagram generation. Comprehensive experiments, including ablations, subjective and objective evaluations, and tests across user expertise levels, demonstrate the effectiveness of our proposed framework. Nous achieves leading efficiency and output quality, while remaining robust to varying user expertise. Moreover, its design is domain-agnostic, and we show evidence of generalization beyond diagram generation. Experimental results prove that our work offers a principled, scalable, and adaptive paradigm for resolving uncertainty about user intent in complex human-AI collaboration.

1 INTRODUCTION

The transition of AI from an efficient tool to a true collaborative partner hinges on solving a core challenge: achieving a shared understanding with the user (Liang & Banks, 2025). While Large Language Models (LLMs) demonstrate remarkable fluency in text generation, their passive, instruction-following nature falters when faced with the inherent incompleteness of human intent expression (Shneiderman, 2022). This limitation is especially evident in creative and technical domains (Amershi et al., 2019; Fan et al., 2023). In such settings, users may hold highly innovative ideas yet struggle to articulate them with precision (Chang et al., 2025). When attempting to realize these ideas with AI, they often fall into a frustrating "guessing game," which in turn forces task goals to emerge gradually and be refined through collaborative processes (Oihane et al., 2024). The gap between a user's high-dimensional mental model and their ability to convey it in a machine-readable format has been described as the "intention gap," (Vanessa et al., 2024) which forces collaboration into inefficient trial-and-error loops (Buccinca, 2020). As a result, the entire burden of precise articulation falls on the human, and this paradigm is fundamentally unsustainable for complex tasks.

Our research stems from a core insight: Why must humans always painstakingly teach the AI, instead of the AI intelligently guiding the human? We advocate for a paradigm shift: envisioning AI not as a passive follower, but as an agent actively bridging this gap (McGrath et al., 2024; Haase & Pokutta, 2024). Inspired by the Socratic method, we treat it not merely as pedagogy but as a model for collaborative discovery (Liu et al., 2024). A Socratic agent does not simply await commands; it formulates strategic questions to systematically resolve its uncertainty about the user's goal (Krishna & et al., 2022; Sahu, 2024). Each question-answer turn becomes a deliberate act of information seeking, designed to maximize convergence toward a shared, high-fidelity understanding (Holstein & et al., 2020; Yao et al., 2025; Khorsand & Pourahmadi, 2025; Thomas & Houssineau, 2024).

To this end, we introduce Nous, an agent designed to acquire proficiency in an optimal inquiry policy. The central mechanism of Nous is a training framework grounded in the first principles of information theory (Cover & Thomas, 2006; Wu et al., 2025; Khandelwal et al., 2025).Within this framework, we define the information gain from dialogue as an intrinsic reward, formally equivalent to the reduction of Shannon entropy over possible task specifications. By relying on this objective and computationally tractable signal,Nous avoids dependence on costly human preference annotations or external reward models (Spera & Agrawal, 2025; Lee & et al., 2022; Agarwal et al., 2022).

To validate this framework, we select scientific diagram generation as our testbed, a prototypical instance of the intention gap. The task is both high-dimensional and logically structured, providing objective criteria for evaluation while remaining sufficiently challenging (Basole & Major, 2024; Han et al., 2023). Building on this, we construct an automated simulation pipeline to generate a large-scale, preference-based dataset tailored to this setting (Shao et al., 2024). Finally, we conducted comprehensive experiments and evaluations, which demonstrated the effectiveness of our method. Moreover, the framework is domain-agnostic: we further show evidence of generalization beyond diagram generation through additional experiments in co-creative contexts (Haase & Pokutta, 2024; Singh et al., 2025). (1) **Nous**, an intelligent agent that instantiates the Socratic interaction paradigm with structured belief modeling. (2) **An information-theoretic reinforcement learning framework**, using dialogue-driven information gain as an intrinsic reward and eliminating the need for human annotation or external reward models. (3) **An automated large-scale simulation pipeline**, generating dialogue strategy learning data to support scalable training and evaluation.

2 Related Work

Our work is situated at the intersection of three key areas in AI and human-computer interaction: goal-oriented dialogue, active learning, and large language model alignment.

Goal-Oriented Dialogue Systems. Traditional goal-oriented dialogue systems, designed for explicit slot-filling tasks like booking flights, excel in closed domains but struggle with the ambiguity of creative and technical tasks (Young et al., 2013; Wen et al., 2017; Budzianowski et al., 2018). Recent work has explored making LLMs more proactive, for example by asking clarification questions in open-domain QA (Rao et al., 2023; Darji & Lutellier, 2025; Wang et al., 2024), by modeling when to inquire based on future dialogue turns (Xu et al., 2024), or by adopting Socratic prompting strategies (Chang, 2023; Zhou et al., 2022). Nevertheless, most LLM-based systems remain passive, relying on the user to drive the interaction. Our work moves beyond this paradigm: rather than filling predefined slots, Nous navigates a combinatorially complex specification space, managing dialogue to resolve uncertainty and transforming the agent from a passive recipient into an active inquirer.

Active Learning and Optimal Experiment Design. The principle of reducing uncertainty by asking questions is rooted in active learning and optimal experiment design (Beluch et al., 2018; Lewis & Gale, 1994). Prior dialogue-policy research has incorporated entropy reduction as a signal for clarification (Padmakumar & Mooney, 2020), and recent studies formalize question quality directly via expected information gain (Mazzaccara et al., 2024; Geishauser et al., 2021; Xing et al., 2024). However, these methods typically target static datasets or constrained "20-questions" benchmarks. Our contribution is to extend this principle to dynamic dialogue for creative design: instead of selecting a data point, Nous learns to generate natural language questions that probe a latent goal space. Training this generative policy with entropy reduction as a real-time reward bridges classical theory with modern LLM interaction (Piriyakulkij et al., 2024; Chen et al., 2025; Zhao et al., 2025).

LLM Alignment and Preference-Based Learning. Aligning LLMs with human intent is a central challenge. Preference-based methods such as RLHF (Christiano et al., 2017; Ouyang & et al., 2022), PPO-based optimization (Schulman et al., 2017), and more recent approaches like GRPO (Shao et al., 2024), DPO (Rafailov et al., 2023), and RLAIF (Bai & et al., 2022; Lee et al., 2024) rely on costly preference labels or heuristic feedback. Our method offers a scalable alternative: we define an intrinsic reward from information gain, bypassing external reward models and the associated annotation cost. By applying offline RL (Levine et al., 2020; Kostrikov & et al., 2022) on automatically generated preferences, Nous avoids proxy misspecification while maintaining principled grounding in task structure, offering a complementary path for alignment in structured co-creative tasks.

AI for Design and Creativity. A growing body of work envisions AI as a co-creative partner in domains such as design and engineering (Tang et al., 2024; Singh et al., 2025). Most systems em-

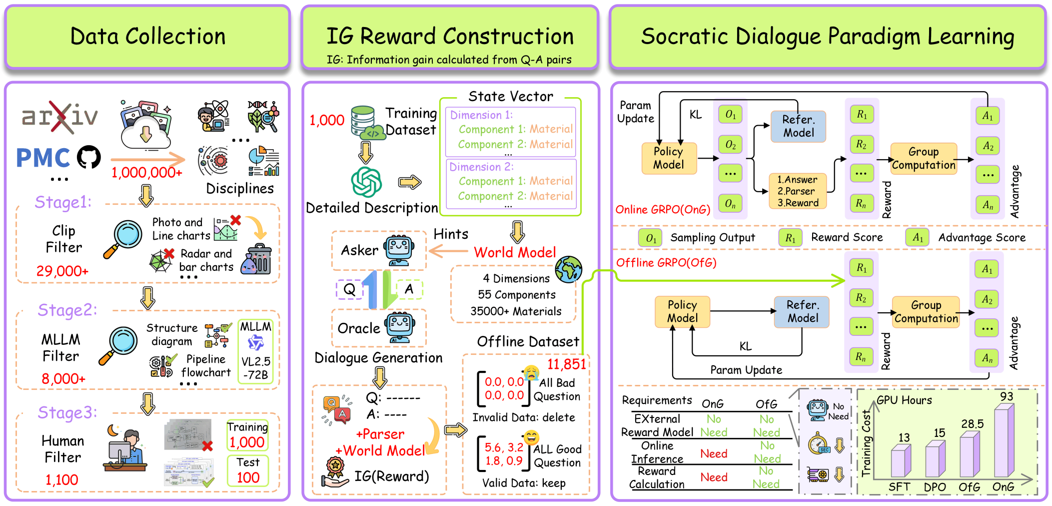

Figure 1: **The multi-stage curation pipeline for the dataset and the details of model training.** We began with a raw dataset of approximately 1 million figures downloaded from scientific papers in different fields on arXiv and PMC. This dataset was first filtered using the CLIP model to remove data plots (such as bar charts and line graphs), resulting in 29,000 images. Next, we used the Qwen-2.5-VL-72B model to retain true schematic diagrams, reducing the dataset to 8,000 images. Finally, three PhD students conducted a manual review to ensure the relevance, clarity, and quality of each figure, resulting in a final dataset of 1,100 images. From this curated dataset, 1,000 figures were used to build the world model and train simulations, while 100 figures were set aside for testing. Detailed explanations regarding data distribution and open-source licenses are provided in Appendix .

phasize generation-providing suggestions or auto-completions. However, effective creation depends on a well-defined goal. Our approach is unique in focusing on the "front-end" of co-creation: clarifying the user's initial, ambiguous intent through dialogue. This emphasis on intent understanding complements existing generative systems and lays a stronger foundation for accurate, relevant, and user-aligned downstream outputs.

3 METHODOLOGY

Our methodology is presented in three parts. First, we establish a formal information-theoretic framework, deriving an intrinsic and tractable reward signal from first principles (Sec. 3.1). Next, we detail the complete offline training pipeline, which includes an automated simulation for preference data generation and the offline policy optimization algorithm. (Sec. 3.2). Finally, we introduce the baseline models used for our comparative experiments (Sec. 3.3).

3.1 AN INFORMATION-THEORETIC FRAMEWORK FOR OPTIMAL INQUIRY

To learn an effective inquiry strategy, the agent requires a quantitative metric for guidance. Drawing from classical information theory, we define a reward signal based on information gain, which measures the informational value of each question-answer turn. We validate our method on the scientific chart generation task, where the dialogue is modeled as a process of reducing epistemic uncertainty over a structured state space. The information gain from a user's response is formally defined as the Kullback-Leibler (KL) divergence between the posterior and prior belief states(the agent's probability distribution over user intentions). We prove this metric simplifies to the reduction in the system's Shannon entropy. This provides an intrinsic reward signal, directly calculable from the agent's belief state, for optimizing the inquiry policy without requiring a separate, pre-trained reward model.

Formalizing the Diagram Specification Space. We begin by defining the object of our inquiry. A complete scientific diagram specification, denoted by $\mathcal{G}$, is conceptualized as a point in a high-dimensional, discrete state space. A diagram specification is represented by a set of N attributes, $\mathcal{V} = \{V_1, V_2, \ldots, V_N\}$. Each attribute V_i takes a value v_i from its finite, discrete domain $\mathcal{S}_i$. A complete and valid diagram specification is an instantiation $\mathbf{g} = (v_1, v_2, \ldots, v_N)$ where $v_i \in \mathcal{S}_i$ for all

$i \in \{1, \ldots, N\}$. The attributes are designed to be comprehensive, covering aspects such as overall layout (V_{layout}), color palettes (V_{color}), the number and types of components ($V_{\text{num\_comp}}, V^{(k)}_{\text{comp\_type}}$), and interconnections ($V^{(i,j)}_{\text{conn}}$).

Quantifying and Decomposing Epistemic Uncertainty. At any turn t in the dialogue, the agent's knowledge about the user's desired diagram is captured by a probabilistic belief state, $P_t(\mathcal{G})$. For computational tractability, we assume the attributes V_i are conditionally independent given the dialogue history $\mathcal{H}_t$. While this is a simplifying assumption, we argue it is a tractable and effective first-order approximation, as the greatest reduction in uncertainty, particularly in early dialogue, comes from resolving major, orthogonal attributes (e.g., overall layout, number of components).

This allows the joint distribution to be factorized:

$$P_t(\mathcal{G}) = P(V_1, \ldots, V_N \mid \mathcal{H}_t) = \prod_{i=1}^{N} P(V_i \mid \mathcal{H}_t). \tag{1}$$

The agent's initial belief state, $P_0(\mathcal{G})$, is an empirical prior derived from a large-scale corpus $\mathcal{D}$ of existing diagrams, where each prior probability is estimated via maximum likelihood:

$$P_0(V_i = v_j) = \frac{\text{Count}_{\mathcal{D}}(V_i = v_j)}{|\mathcal{D}|}. \tag{2}$$

The total uncertainty of the system is the Shannon entropy of the belief state $P_t(\mathcal{G})$. A critical consequence of the independence assumption is that the total entropy decomposes into a sum of marginal entropies:

$$\mathcal{H}(P_t(\mathcal{G})) = -\sum_{\mathbf{g} \in \mathcal{G}} P_t(\mathbf{g}) \log_2 P_t(\mathbf{g}) = \sum_{i=1}^{N} \mathcal{H}(P_t(V_i))\,, \tag{3}$$

where $\mathcal{H}(P_t(V_i)) = -\sum_{v_j \in \mathcal{S}_i} P_t(V_i = v_j)\, \log_2 P_t(V_i = v_j)$. This decomposition is crucial, as it allows us to track uncertainty on a per-attribute basis.

Belief State Update and Reward Function. The dialogue proceeds as a sequence of belief state updates. An answer A_t is mapped by a semantic parser f to evidence $\mathcal{E}_t$, which imposes hard constraints on a subset of attributes $\mathcal{V}_{\mathcal{E}_t}$. In our simulation, f is implemented as a few-shot prompted LLM, whose parsing accuracy is ensured by the Oracle's templated responses, providing a reliable signal for reward calculation. This updates the belief from a prior P_t to a posterior P_{t+1} via Bayesian conditioning. For any constrained attribute, the posterior becomes a deterministic Kronecker delta function, $P_{t+1}(V_i = v_j) = \delta_{jk}$, while unconstrained attributes remain unchanged.

We define our reward signal r_t as the *reduction in Shannon entropy* of the belief state:

$$r_t \equiv IG(A_t) \;=\; \mathcal{H}(P_t(\mathcal{G})) - \mathcal{H}(P_{t+1}(\mathcal{G}))\,. \tag{4}$$

Intuitively, this quantity measures the informational value of the user's answer. From an information-theoretic perspective, the expected value of this entropy reduction equals the mutual information between A_t and $\mathcal{G}$, which can be written as an expectation over a KL divergence:

$$\mathbb{E}[IG(A_t)] \;=\; I(A_t; \mathcal{G}) \;=\; \mathbb{E}_{A_t}\big[\, D_{KL}\big(P_{t+1}(\mathcal{G}) \,\|\, P_t(\mathcal{G})\big)\big]\,. \tag{5}$$

Thus maximizing information gain is identical to maximizing the reduction of uncertainty.

By substituting the entropy decomposition from Eq. 3 into Eq. 4, we derive a tractable reward function:

$$r_t = \left(\sum_{i=1}^{N} \mathcal{H}(P_t(V_i))\right) - \left(\sum_{i=1}^{N} \mathcal{H}(P_{t+1}(V_i))\right) \;=\; \sum_{i=1}^{N} \left(\mathcal{H}(P_t(V_i)) - \mathcal{H}(P_{t+1}(V_i))\right). \tag{6}$$

Under our hard-constraint update model, the posterior entropy $\mathcal{H}(P_{t+1}(V_i))$ becomes zero for any newly constrained attribute $V_i \in \mathcal{V}_{\mathcal{E}_t}$, and remains unchanged for all other attributes. Therefore, the sum in Eq. 6 simplifies to include only the terms for the resolved attributes:

$$r_t \;=\; \sum_{V_i \in \mathcal{V}_{\mathcal{E}_t}} \mathcal{H}(P_t(V_i))\,. \tag{7}$$

This final equation states that the utility of an answer is the sum of the prior entropies of the attributes it clarifies. This signal is intrinsic, computationally efficient, and provides a robust foundation for optimizing the agent's inquiry policy.

3.2 OFFLINE POLICY OPTIMIZATION

With a defined reward signal, we can now train the agent's inquiry policy. Our approach is a fully offline process, which enhances stability and computational efficiency. It consists of two main stages: first, we generate a large-scale, static dataset of preference-ranked inquiries through simulation; second, we use this dataset to train the policy via an offline reinforcement learning algorithm.

Automated Preference Data Generation Our training process relies on a large-scale preference dataset, $\mathcal{D}_{\text{pref}}$, which we generate through an automated simulation framework. This simulation requires two key components: a "world model" to provide prior probabilities (as in Eq. 2) and a set of ground-truth tasks. We construct both from a high-quality corpus of scientific diagrams, curated through a multi-stage filtering pipeline detailed in Figure 1.

Within the simulation, an "Oracle" agent, holding a ground-truth specification from our curated set, provides answers to inquiries proposed by multiple candidate models. The information gain for each inquiry is calculated via Eq. 7, yielding a training sample $\{p, \{r_1, \ldots, r_k\}, \{R_1, \ldots, R_k\}\}$, consisting of a prompt, a group of candidate responses, and their corresponding reward scores.

Offline Adaptation of Group Relative Policy Optimization. To optimize our policy π_θ on the static dataset $\mathcal{D}_{\text{pref}}$, we adapt the objective function from Group Relative Policy Optimization (GRPO) for an offline setting. While GRPO was originally proposed as an online algorithm that iteratively samples from the policy, we find its core objective is well-suited for offline training in our context. The rationale for this offline adaptation is twofold. First, the task of "asking a good question" is a capability already inherent in pretrained LLMs. The distribution of our generated candidate responses is therefore not expected to be drastically different from what the policy would generate, making on-policy sampling less critical. Second, using a static dataset eliminates the computational overhead of online generation, leading to a much more efficient and stable training pipeline.

For each group of responses, we first normalize the rewards into advantage estimates $A(r_i, p)$ via z-scoring within the group. This stabilizes the learning process across different prompts. Our offline algorithm then maximizes the following PPO-style clipped surrogate objective:

$$L_{\text{Policy}}(\theta) = \mathbb{E}_{(p,r_i,A_i)\sim\mathcal{D}_{\text{pref}}}\left[\min\left(\rho_i(\theta)A_i, \text{clip}(\rho_i(\theta), 1-\epsilon, 1+\epsilon)A_i\right)\right] \tag{8}$$

where the probability ratio $\rho_i(\theta) = \pi_\theta(r_i|p)/\pi_{ref}(r_i|p)$measures the policy change against a frozen reference policy π_{ref}. The clipping function clip$(\cdot)$ constrains this ratio to a trusted region, preventing overly aggressive and destabilizing policy updates.

To further regularize the policy and ensure it does not deviate excessively from the pre-trained base model, we incorporate a KL-divergence penalty, leading to the final loss function:

$$L_{\text{total}}(\theta) = L_{\text{Policy}}(\theta) - \beta D_{KL}(\pi_\theta(\cdot|p) \,||\, \pi_{ref}(\cdot|p)) \tag{9}$$

where β is a hyperparameter controlling the strength of the KL penalty. The log-probabilities $\log \pi(r|p)$ are computed autoregressively. To ensure the policy is only trained on its generation, we apply a loss mask so that the gradients are backpropagated only through the tokens of the response r, not the prompt p.

3.3 CONTRASTING METHODS FOR ABLATION STUDY

To rigorously evaluate the effectiveness of the **offline GRPO (OfG)** paradigm, we will use several other key baselines to train Nous for comparison in the experiments.

Supervised Fine-Tuning (SFT): A baseline model fine-tuned only on the highest-reward (prompt, response) pairs from our dataset. This helps isolate the contribution of preference-based optimization over simple imitation learning. **Direct Preference Optimization (DPO)**: To compare against a prominent pairwise preference learning method, we implement a DPO baseline. DPO optimizes the policy to directly increase the log-probability ratio of preferred to dispreferred responses, using only the best and worst responses from each group. **Online GRPO (OnG)**: To validate the efficiency and stability of the offline approach, we also train a model using an online GRPO pipeline. This involves an initial SFT warm-up, followed by an iterative process of sampling responses from the policy, calculating their rewards, and updating the policy. All training methods ultimately include an SFT to train their ability for final integrated description.

4 Experiments

We conduct a comprehensive set of experiments to evaluate our proposed framework. Our evaluation is designed to answer four key research questions: (1) Does our information-theoretic approach lead to more efficient interactions compared to established baselines? (2) Does higher interaction efficiency translate to superior quality in the final generated artifact? (3) Is the information gain-based reward signal the critical component of our framework's success? (4) How robust is the learned inquiry policy to variations in user expertise?

4.1 Experimental Setup

Models Under Evaluation. Our primary model, Nous, is built upon Qwen3-8B and trained with full-parameter fine-tuning. For evaluation, we consider two groups of baselines. Trained Baselines: three Nous variants trained with alternative methods (SFT, DPO, OnG; see Section 3.3). Prompt-Based Baselines: a proprietary model (GPT-5: GPT-few ,GPT-zero) and a large open-source model (Qwen3-235B: Qwen-few, Qwen-zero), each tested under zero-shot and few-shot prompting. All prompts are instantiated using the *Socratic prompting* paradigm (Chang, 2023), which encourages the model to ask clarifying questions before producing a figure. We include these as the most relevant horizontal comparison, since no other mature baselines exist for scientific figure generation. Full prompt templates and hyperparameters are given in the Appendix.

Evaluation Task and Data. We take the task of scientific diagram generation in human-AI collaboration as our test scenario. The test data comes from a hold-out set of 100 complex real-world diagrams (see Figure1, for detailed sources see Appendix E). For each diagram, we simulate an interaction where the agent must elicit the complete specification from an Oracle. The Oracle, which holds the ground-truth specification for a target diagram and is configured identically to the one used for generating our training data. Each dialogue begins with a generic initial request, "I want to create a scientific diagram," and concludes when the agent indicates it has gathered sufficient information by outputting a final, consolidated description of the diagram. This automated simulation ensures a fair, controlled, and reproducible comparison across all models.

Evaluation Metrics. We employ a multifaceted evaluation strategy to assess both the process and the outcome: **Interaction Efficiency**: (1)We measure this by the average number of turns an agent takes to complete the dialogue, (2)and the cumulative information gain achieved throughout the interaction. Higher efficiency is indicated by fewer turns and a steeper information gain curve. **Output Quality**: We assess the quality of the final specification from two complementary angles: (1) subjective preference scores, where the final generated diagrams are evaluated by human and AI judges through pairwise comparisons, and (2) a suite of objective, specification-based metrics that quantitatively score the generated diagrams against the ground truth.

4.2 Main Results

4.2.1 Interaction Efficiency

Dialogue Turns and Resource Cost: Table 1 details the average number of dialogue turns and the associated training costs. First, all trained models complete the task in fewer turns than their non-trained counterparts. Then the performance of the SFT-trained agent shows the highest dialogue efficiency and the lowest training resources, but this brevity corresponds to the lowest total information gain among all trained models, indicating a premature and incomplete inquiry process. In contrast, the agent trained with OfG maintains a competitive turn count while requiring resources only marginally higher than DPO and significantly lower than OnG. This result highlights the scalability and cost-effectiveness of our offline training framework.

Information Gain (IG) Dynamics: Figure 2 plots the cumulative information gain against the number of dialogue turns, offering a more granular view of the inquiry strategies. The agents trained via OnG and OfG exhibit the most sustained information gain curves, demonstrating a robust ability to consistently pose high-value questions throughout the interaction. The SFT-trained agent, however, reveals a critical weakness: despite a strong start, its performance mirrors that of the non-trained models after the initial turns. They all fall into an "information bottleneck," where the ability to ask meaningful, probing questions sharply diminishes, causing their gain curves to flatten. This empirically validates the "frustrating guessing game" that motivated our work and underscores the necessity of a structured, goal-oriented training paradigm to overcome this fundamental limitation.

Table 1: Experimental results of interaction efficiency, training resource consumption, and dynamic information gain.

Model	Turns (↓)	Total IG (↑)	Resource hours(↓)	Information Gain Dynamics at Turn (↑) Turn 1	Turn 5	Turn 10	Turn 15	Turn 20
Nous (OfG)	20.3	**120.5**	28.5	10.4	66.6	**99.1**	**113.7**	**120.5**
Nous (OnG)	22.0	115.8	93	7.8	59.4	88.4	107.2	114.3
Nous (DPO)	21.5	111.3	15	**13.9**	65.8	90.7	101.5	110.9
Nous (SFT)	**17.1**	94.3	**13**	12.6	**78.1**	90.5	94	94.3
GPT-few	22.6	93.5	N/A	9.1	60.4	77.4	88.1	92.1
GPT-zero	26.5	84.8	N/A	11.3	43.2	59.7	72.7	78.3
Qwen-few	19.5	90.5	N/A	10.6	61.1	76.4	85.9	90.5
Qwen-zero	25.3	81.5	N/A	6.6	48.1	64.4	77.2	80.3

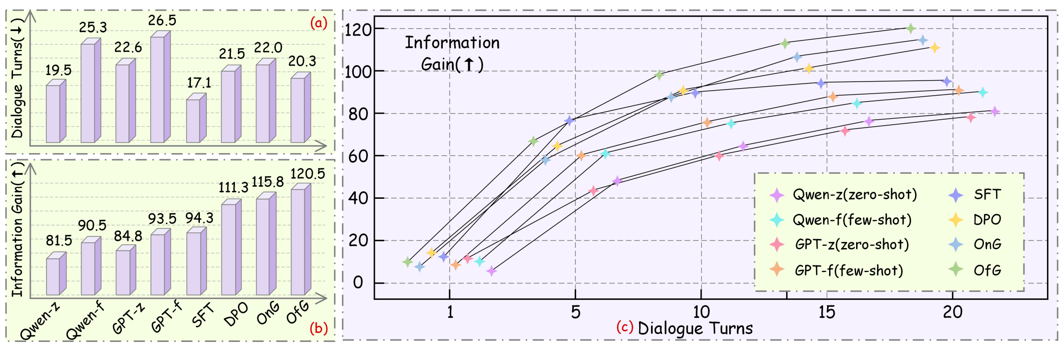

Figure 2: Experimental results of Interaction Efficiency. (a) The average number of dialogue turns for each model to complete information collection; (b) The average information gain obtained during the dialogue for each model; (c) The dynamic change of information gain during the dialogue

4.2.2 Output Quality

Subjective comparison We used two text-to-image backbone models (4o-image-1 and nano-banana) to generate two images based on the final natural language specifications of each model. These images were evaluated twice for their drawing quality through pairwise comparisons by human reviewers and GPT-5 with reference to the test set, resulting in a total of 11,200 comparisons. Table 2 reports the results under the evaluation protocols of three tie-handling methods, the results from the two judging protocols were highly consistent, lending reliability to the evaluation setup. Among the trained models, Nous trained with OfG and OnG achieved the highest win rates, outperforming DPO and SFT. The non-trained baselines lagged behind, with GPT-based models generally stronger than Qwen-based ones. Detailed pairwise results are visualized in Figure 3, and case studies are provided in the Appendix J.

Objective metrics: To complement these subjective judgments with reproducible quantitative scores, we employed the VisPainter framework,a tool that converts text specifications into editable vector graphics, with examples and descriptions provided in the Appendix E and J. This evaluates diagram specifications across six dimensions: Precision, Recall, Design Error Rate, Blank Ratio, Readability, and Alignment. Weighted score is calculated by applying weights of [0.2, 0.2, 0.2, 0.05, 0.25, 0.1] to these six dimensions, shown in Table 3 and Figure 4(a), highlight clear differences: OnG and OfG perform better in terms of drawing precision, recall, and readability. This is attributed to more detailed and information-rich image descriptions. The same applies to the blank ratio; thanks to more abundant component information, more efficient space utilization is achieved. Unexpected results were observed in terms of design error rate and alignment. This is because the number of output elements is proportional to the chance of making mistakes during the drawing process, so SFT and prompt-based baseline models show higher scores in error rate and alignment. These patterns further confirm that models trained with principled inquiry signals have advantages over untrained models.

Table 2: Model win rate results under different tie-handling protocols: (1) “Win”: ties not counted; (2) “W/T(0.5)”: ties contribute 0.5; (3) “W/T”: ties count as 1. All win-rate proportions are based on 400 pairwise judgments per model pair (100 prompts $\times$ 2 judges $\times$ 2 renderers); the standard error of a proportion is at most 0.025, so all 95% confidence intervals are within $\pm$0.05.

Model	4o-image-1						nano-banana					
	Human Judge(↑)			GPT-5 Judge(↑)			Human Judge(↑)			GPT-5 Judge(↑)		
	Win	**W/T(0.5)**	**W/T**	**Win**	**W/T(0.5)**	**W/T**	**Win**	**W/T(0.5)**	**W/T**	**Win**	**W/T(0.5)**	**W/T**
Nous (OfG)	**0.68**	0.71	0.73	**0.69**	**0.72**	**0.76**	**0.61**	**0.66**	**0.72**	**0.55**	0.61	0.66
Nous (OnG)	0.69	**0.72**	**0.75**	0.63	0.67	0.71	0.60	0.65	0.70	0.54	**0.61**	**0.67**
Nous (DPO)	0.59	0.61	0.64	0.56	0.61	0.67	0.57	0.64	0.71	0.54	0.59	0.65
Nous (SFT)	0.49	0.51	0.53	0.42	0.48	0.55	0.38	0.48	0.57	0.37	0.48	0.59
GPT-few	0.45	0.47	0.50	0.45	0.52	0.58	0.34	0.44	0.54	0.39	0.47	0.56
GPT-zero	0.29	0.32	0.35	0.27	0.32	0.37	0.29	0.37	0.46	0.36	0.47	0.57
Qwen-few	0.36	0.40	0.43	0.35	0.42	0.48	0.35	0.42	0.49	0.32	0.41	0.49
Qwen-zero	0.23	0.27	0.30	0.24	0.28	0.32	0.26	0.33	0.40	0.28	0.36	0.44

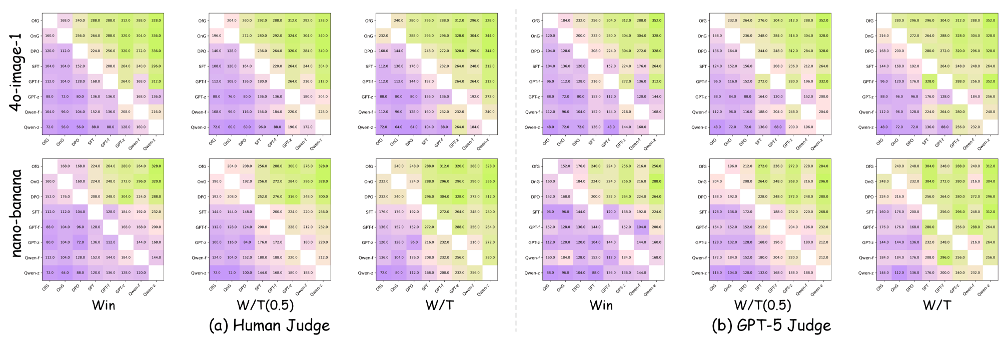

Figure 3: Model scores under different tie-handling protocols. (a) Results of human evaluation; (b) Results of GPT-5 model evaluation.

4.3 Ablation Studies

Reward Function: To validate the critical role of our proposed information-theoretic reward signal, we conducted an ablation study. We replaced it with a heuristic-based “slot-counting” reward, which simply counts the number of specified attributes in each turn and treats all attributes equally. Using this new reward, we trained a model variant named Nous-Counting with the same OfG method on a dataset of identical scale, generated using the process from Section 3.2.

We evaluated this model under the identical experimental setup, with the results presented in Table 4 and 4(b). Nous-Counting completes dialogues in fewer turns, but it achieves substantially lower information gain and final output quality. This is because the slot-counting reward encourages a greedy policy that maximizes the quantity of resolved attributes, not their informational value. The model learns to ask broad, low-impact questions rather than strategically targeting high-entropy attributes first. This study confirms that our information-theoretic reward is essential for guiding the agent to learn an inquiry strategy that is not just superficially fast, but deeply effective.

User Expertise: Real world collaboration involves users with diverse levels of expertise. We evaluated the robustness of Nous (Nous-Entropy) by testing it against three user personas: an Expert Oracle that uses precise technical terms (e.g., “directed acyclic graph”), a Novice Oracle that uses vague, descriptive language (e.g., “show it like a flowchart... with no loops”), and a group of three doctoral students as Human Users who provided descriptions of real usage scenarios.

As shown in Table 4 and Figure 4(c), Nous demonstrates strong adaptability across all user types. In terms of interaction turns, the Novice Oracle required more rounds to resolve ambiguity, whereas the Human users tended to disclose more information per turn, resulting in slightly fewer turns overall. Nevertheless, the final output quality, measured by both subjective and objective scores, showed no

Table 3: Results of the final generated charts using the VisPainter framework. Higher scores in each item are better, and the design error rate has also been inverted to follow the same principle.

Model	Precision	Recall	Design	Blank	Readability	Alignment	Score
Nous (OfG)	0.83	0.84	0.51	**0.83**	0.79	0.88	0.76
Nous (OnG)	**0.84**	**0.86**	0.49	0.81	**0.80**	0.90	**0.77**
Nous (DPO)	0.80	0.81	0.52	0.78	0.75	0.87	0.74
Nous (SFT)	0.76	0.79	0.53	0.74	0.71	0.91	0.72
GPT-few	0.63	0.74	0.51	0.69	0.59	0.93	0.65
GPT-zero	0.42	0.77	**0.55**	0.61	0.41	0.93	0.57
Qwen-few	0.67	0.73	0.53	0.66	0.64	0.91	0.67
Qwen-zero	0.40	0.78	0.54	0.67	0.38	**0.93**	0.57

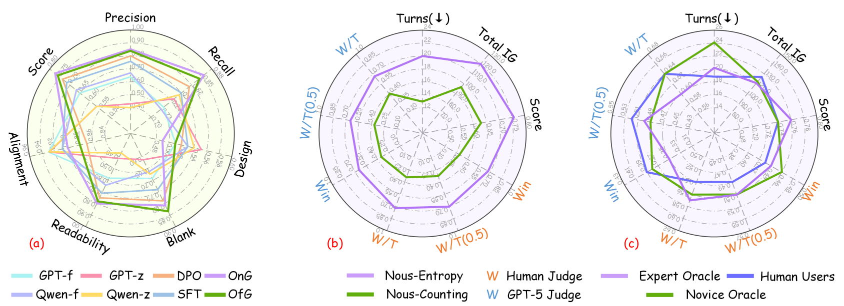

Figure 4: Visualization of experimental results. (a) Evaluation results of each model; (b) Results of ablation experiment 1; (c) Results of ablation experiment 2.

Table 4: Experimental results of the ablation study. Reward function ablation experiment (upper part): Nous-Counting and Nous-Entropy are models trained based on the counting reward and the information entropy reward, respectively; Professional level ablation experiment (lower part). Conduct model image generation evaluation experiments using nano-banana.

	Turns	Total IG	Score	Human Judge(↑)			GPT-5 Judge(↑)		
Method	(↓)	(↑)	(↑)	Win	W/T(0.5)	W/T	Win	W/T(0.5)	W/T
Nous-Entropy	20.3	**120.53**	**0.76**	**0.68**	**0.70**	**0.72**	**0.60**	**0.63**	**0.66**
Nous-Counting	**13.6**	97.11	0.63	0.28	0.30	0.32	0.34	0.37	0.40
Expert Oracle	20.3	120.53	**0.76**	0.44	**0.51**	**0.57**	0.35	0.50	0.64
Novice Oracle	24.1	122.47	0.74	**0.46**	0.51	0.56	0.38	0.49	0.60
Human User	**18.7**	**128.01**	0.74	0.43	0.49	0.54	**0.39**	**0.52**	**0.64**

degradation. This highlights a central advantage of our Socratic framework: its iterative inquiry process is inherently designed to resolve ambiguity. Rather than relying on flawless user input, it strategically poses follow-up questions to progressively converge on the user's intent, demonstrating its effectiveness as a collaborative partner that can accommodate natural human expression.

5 Conclusion

This paper addresses a bottleneck in human-AI collaboration: "intention expression gap." We shift the paradigm from passive instruction-following to active, Socratic collaboration, introducing Nous, an agent that learns to resolve uncertainty about user intent through thoughtful inquiry. Our contribution is a training framework grounded in information theory, defining information gain as an intrinsic reward to eliminate costly human annotation and external reward models. We further show that Offline GRPO provides an efficient and stable path for training such agents. Experiments demonstrate that Nous achieves leading efficiency and output quality, while ablations confirm that the information-theoretic reward, rather than simple heuristics, is the decisive factor, and the agent remains robust across diverse levels of user expertise. In sum, this work presents a principled, scalable, and adaptive paradigm for resolving intent ambiguity, shifting the communication burden away from humans and moving us closer to AI partners capable of genuine collaborative thought.

Reproducibility Statement The models, prompts, data generation code, and model training code we used are all open-source. We have provided the code required to reproduce our research results in the supplementary materials. After the blind review period, we will release the complete code repository. To ensure the reproducibility of this paper, we have made efforts in the following aspects: (1) The code and data will be open-sourced once the paper is accepted. (2) We have conducted extensive experiments under different settings to verify the general applicability of the proposed framework. (3) We have provided a framework and evaluation methods based on open-source models, significantly improving reproducibility.

References

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Reincarnating reinforcement learning: Reusing prior computation to accelerate progress, 2022. URL https://arxiv.org/abs/2206.01626.

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. Guidelines for human-ai interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pp. 1–13, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300233. URL https://doi.org/10.1145/3290605.3300233.

Yuntao Bai and et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Rahul C. Basole and Timothy Major. Generative ai for visualization: Opportunities and challenges. *IEEE Computer Graphics and Applications*, 44(2):55–64, 2024. doi: 10.1109/MCG.2024.3362168.

William H. Beluch, Tim Genewein, A. Nürnberger, and Jan M. Köhler. The power of ensembles for active learning in image classification. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9368–9377, 2018. URL https://api.semanticscholar.org/CorpusID:52838058.

Zana. Buccinca. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *CHI*, 2020.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *EMNLP*, 2018.

Edward Y. Chang. Prompting large language models with the socratic method, 2023. URL https://arxiv.org/abs/2303.08769.

Zongyu Chang, Feihong Lu, Ziqin Zhu, Qian Li, Cheng Ji, Zhuo Chen, Hao Peng, Yang Liu, Ruifeng Xu, Yangqiu Song, Shangguang Wang, and Jianxin Li. Bridging the gap between llms and human intentions: Progresses and challenges in instruction understanding, intention reasoning, and reliable generation, 2025. URL https://arxiv.org/abs/2502.09101.

Minghan Chen, Guikun Chen, Wenguan Wang, and Yi Yang. Seed-grpo: Semantic entropy enhanced grpo for uncertainty-aware policy optimization. *arXiv preprint arXiv:2505.12346*, 2025. URL https://arxiv.org/abs/2505.12346.

Paul Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *NeurIPS*, 2017.

Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. Wiley, 2006.

Harsh Darji and Thibaud Lutellier. Curiosity by design: An llm-based coding assistant asking clarification questions, 2025. URL https://arxiv.org/abs/2507.21285.

Caoyun Fan, Jindou Chen, Yaohui Jin, and Hao He. Can large language models serve as rational players in game theory? a systematic analysis, 2023. URL https://arxiv.org/abs/2312.05488.

Christian Geishauser, Songbo Hu, Hsien-chin Lin, Nurul Lubis, Michael Heck, Shutong Feng, Carel van Niekerk, and Milica Gašić. What does the user want? information gain for hierarchical dialogue policy optimisation. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 2021. URL https://arxiv.org/abs/2109.07129.

Jennifer Haase and Sebastian Pokutta. Human-ai co-creativity: Exploring synergies across levels of creative collaboration, 2024. URL https://arxiv.org/abs/2411.12527.

Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. Chartllama: A multimodal llm for chart understanding and generation, 2023. URL https://arxiv.org/abs/2311.16483.

Kenneth Holstein and et al. Participatory approaches to human-ai alignment. In *CHI*, 2020.

Zhiyuan Hu, Chumin Liu, Xidong Feng, Yilun Zhao, See-Kiong Ng, Anh Tuan Luu, Junxian He, Pang Wei Koh, and Bryan Hooi. Uncertainty of thoughts: Uncertainty-aware planning enhances information seeking in large language models, 2024. URL https://arxiv.org/abs/2402.03271.

Eleni Ilkou, Stephan Linzbach, and Jonas Wallat. Hybrid evaluation of socratic dialogue for teaching. *CEUR Workshop Proceedings*, 3953, 2025.

Anant Khandelwal, Manish Gupta, and Puneet Agrawal. Cocoa: Confidence and context-aware adaptive decoding for resolving knowledge conflicts in large language models, 2025. URL https://arxiv.org/abs/2508.17670.

Hadi Khorsand and Vahid Pourahmadi. Ofal: An oracle-free active learning framework, 2025. URL https://arxiv.org/abs/2508.08126.

Ilya Kostrikov and et al. Offline reinforcement learning with implicit q-learning. In *ICLR*, 2022.

Ranjay Krishna and et al. Visual question generation for interactive ai. In *CVPR*, 2022.

Daniel Lee, Sungmin Park, and Kyunghyun Cho. Rlaif: Scaling reinforcement learning from ai feedback. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. OpenReview preprint.

Mina Lee and et al. A survey on simulation for human-ai interaction. *Foundations and Trends in HCI*, 2022.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

David D Lewis and William A Gale. A sequential algorithm for training text classifiers. *SIGIR*, 1994.

Qingyu Liang and Jaime Banks. On the same page: Dimensions of perceived shared understanding in human-ai interaction, 2025. URL https://arxiv.org/abs/2505.20068.

Jiayu Liu, Zhenya Huang, Tong Xiao, Jing Sha, Jinze Wu, Qi Liu, Shijin Wang, and Enhong Chen. Socraticlm: exploring socratic personalized teaching with large language models. *Advances in Neural Information Processing Systems*, 37:85693–85721, 2024.

Davide Mazzaccara, Alberto Testoni, and Raffaella Bernardi. Learning to ask informative questions: Enhancing llms with preference optimization and expected information gain. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024. URL https://aclanthology.org/2024.findings-emnlp.291/.

Melanie J. McGrath, Andreas Duenser, Justine Lacey, and Cecile Paris. Collaborative human-ai trust (chai-t): A process framework for active management of trust in human-ai collaboration, 2024. URL https://arxiv.org/abs/2404.01615.

Oihane, Diego Casado-Mansilla, Diego López de Ipiña, and Javier García-Zubia. Human-in-the-loop machine learning: Reconceptualizing the role of the user in interactive approaches. *Internet of Things*, 25:101048, 2024. ISSN 2542-6605. doi: https://doi.org/10.1016/j.iot.2023.101048. URL https://www.sciencedirect.com/science/article/pii/S2542660523003712.

Long Ouyang and et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

Aishwarya Padmakumar and Raymond J. Mooney. Dialog policy learning for joint clarification and active learning queries, 2020. URL https://arxiv.org/abs/2006.05456.

Wasu Top Piriyakulkij, Volodymyr Kuleshov, and Kevin Ellis. Active preference inference using language models and probabilistic reasoning, 2024. URL https://arxiv.org/abs/2312.12009.

Rafael Rafailov et al. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2023.

Sudha Rao, Lei Zhang, Dian Yu, and et al. Asking clarification questions to handle ambiguity in open-domain qa. In *Findings of EMNLP*, 2023.

Abinash Sahu. Information gain, operator spreading, and sensitivity to perturbations as quantifiers of chaos in quantum systems, 2024. URL https://arxiv.org/abs/2404.09464.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. In *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Ben Shneiderman. *Human-Centered AI*. Oxford University Press, 01 2022. ISBN 9780192845290. doi: 10.1093/oso/9780192845290.001.0001. URL https://doi.org/10.1093/oso/9780192845290.001.0001.

Saloni Singh, Koen Hindriks, Dirk Heylen, and Kim Baraka. A systematic review of human-ai co-creativity, 2025. URL https://arxiv.org/abs/2506.21333.

Cosimo Spera and Garima Agrawal. Reversing the paradigm: Building ai-first systems with human guidance, 2025. URL https://arxiv.org/abs/2506.12245.

Ryuichi Takanobu, Hanlin Zhu, and Minlie Huang. Guided dialog policy learning: Reward estimation for multi-domain task-oriented dialog. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 100–110, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1010. URL https://aclanthology.org/D19-1010/.

Xiangru Tang, Howard Dai, Elizabeth Knight, Fang Wu, Yunyang Li, Tianxiao Li, and Mark Gerstein. A survey of generative ai for de novo drug design: New frontiers in molecule and protein generation, 2024. URL https://arxiv.org/abs/2402.08703.

Jake Thomas and Jeremie Houssineau. Improving active learning with a bayesian representation of epistemic uncertainty, 2024. URL https://arxiv.org/abs/2412.08225.

Vanessa, Xiaotong Zhang, and Kamal Youcef-Toumi. Bayesian intention for enhanced human robot collaboration, 2024. URL https://arxiv.org/abs/2410.00302.

Qian Wang, Ming Liu, and Rui Zhao. Boosting large language models with the socratic method for math teaching. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*. ACM, 2024.

Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. A network-based end-to-end trainable task-oriented dialogue system. In *EACL*, 2017.

Fei Wu, Pablo Marquez-Neila, Hedyeh Rafi-Tarii, and Raphael Sznitman. Active learning with context sampling and one-vs-rest entropy for semantic segmentation, 2025. URL `https://arxiv.org/abs/2412.06470`.

Teng Xiao, Zhen Ge, Sujay Sanghavi, Tian Wang, Julian Katz-Samuels, Marc Versage, Qingjun Cui, and Trishul Chilimbi. Infopo: On mutual information maximization for large language model alignment. In *Proceedings of the 2025 NAACL: Human Language Technologies, Long Papers*, pp. 11699–11711, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

Xiaoying Xing, Peixi Xiong, Lei Fan, Yunxuan Li, and Ying Wu. Learning to ask denotative and connotative questions for knowledge-based VQA. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8301–8315, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.487. URL `https://aclanthology.org/2024.findings-emnlp.487/`.

Zhen Xu et al. Modeling future conversation turns to teach llms to ask clarifying questions. In *OpenReview*, 2024.

Shanle Yao, Ghazal Alinezhad Noghre, Armin Danesh Pazho, and Hamed Tabkhi. Alfred: An active learning framework for real-world semi-supervised anomaly detection with adaptive thresholds, 2025. URL `https://arxiv.org/abs/2508.09058`.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179, 2013.

Weixiang Zhao, Xingyu Sui, Yulin Hu, Jiahe Guo, Haixiao Liu, Biye Li, Yanyan Zhao, Bing Qin, and Ting Liu. Teaching language models to evolve with users: Dynamic profile modeling for personalized alignment, 2025. URL `https://arxiv.org/abs/2505.15456`.

Andy Zhou, Jean-Baptiste Alayrac, and et al. Socratic models: Composing zero-shot multimodal reasoning with language. In *ICLR*, 2022.

6 Appendix

A Statement on LLMs Usage

The authors used large language models (LLMs) during the writing process solely for language refinement and editing. It should be explicitly stated that LLMs were not employed in any core aspects of the research, including the formulation of research ideas, the design of methodologies, the execution of experiments, or the development of conclusions. All scholarly contributions were made independently by the authors.

B Extended Discussion of Related Work

Clarification and Inquiry as Strategy. A growing body of work recognizes the strategic value of asking clarifying questions. In open-domain QA, for example, clarification has been shown to improve accuracy by resolving ambiguity before answering (Rao et al., 2023). Other approaches model the decision of whether and when to ask a question based on the expected utility of future dialogue turns, effectively learning an optimal timing policy (Xu et al., 2024). In specialized domains like code generation, clarification also improves correctness, highlighting its broad value (Darji & Lutellier, 2025). While these methods validate the importance of proactive inquiry, they often optimize for single-answer correctness using heuristic signals or rely on downstream annotations to estimate future value. Nous shifts the focus from when to ask to what to ask. Our framework aims for convergence toward a complete, high-dimensional specification, where the reward is an immediate, intrinsic signal derived from entropy reduction over structured attributes, providing a stable, cumulative signal for optimizing the content of each inquiry.

Information Gain as a Measure of Question Quality. Our work builds on the principle of using information theory to quantify question value. In task-oriented dialogue, early frameworks used reward estimation to guide policy learning, though often as a proxy for external goals like booking success (Takanobu et al., 2019; Geishauser et al., 2021). More directly, work in visual dialogue has used information gain to explicitly model the value of "confirmation questions" (e.g., yes/no questions), demonstrating that such inquiries efficiently reduce the candidate set and improve success rates in guessing games like GuessWhat?! (Hu et al., 2024). Similarly, recent research establishes the "20 questions" game as a benchmark for active information seeking in LLMs, using expected information gain to rank and select the most discriminative question from a set of candidates generated via Chain-of-Thought prompting (Mazzaccara et al., 2024; Sahu, 2024). These studies collectively affirm that an entropy-based objective is a powerful tool for guiding efficient inquiry. Nous integrates and advances these ideas into a scalable learning paradigm. Instead of using information gain as a post-hoc selection heuristic (Xiao et al., 2025) or applying it to a constrained set of question types, we use it as a real-time, intrinsic reward to train a generative policy. This enables Nous to learn to generate open-ended, natural language questions, which offers a significant advantage in high-dimensional, structured design spaces. In this way, we bridge the gap between the theoretical appeal of information gain and the practical challenge of training a proactive conversational agent for complex, creative tasks.

Socratic Prompting versus Learnable Strategy. Socratic prompting, exemplified by *Prompting Large Language Models with the Socratic Method* (Chang, 2023), encourages models to ask questions before answering through templates. *SocraticLM: Exploring Socratic Personalized Teaching* (Liu et al., 2024) extends this to personalized instruction, while *Hybrid Evaluation of Socratic Dialogue for Teaching* (Ilkou et al., 2025) evaluates its educational benefits and limits. While these approaches highlight the pedagogical value of Socratic interaction, they remain prompt-based or domain-specific. Nous extends the paradigm into a trainable policy: information gain defines the objective, and offline preference data enables optimization. This transforms "asking questions" from prompt-driven behavior into a generalizable capability robust across user types.

Summary. Prior work can be grouped into two broad directions: Socratic prompting methods that encourage proactive questioning through templates or pedagogy, and a method to achieve accurate question answering by quantifying the value of questions through entropy or mutual information. Nous advances both threads by combining structured belief states, closed-form entropy-based rewards, and offline policy optimization, thereby making clarification a scalable, principled, and generalizable strategy rather than a heuristic or template.

C Supplementary to Ablation Experiments

To investigate the impact of training data quality on the final policy, we adjusted the data generation process. In addition to the Template Oracle, we introduced two variants: a Vague Oracle (providing incomplete information) and a Noisy Oracle (interjecting irrelevant information in its responses). Using these three data sources, we trained three respective models: Nous-Template, Nous-Vague, and Nous-Noisy. Distinct from the discussion on user expertise in Section 4.3, this section evaluates our framework's learning ability under different training data conditions.

The experimental results, shown in Table 5, reveal the following: **Adaptability to Vagueness.** Nous-Vague's performance in standard tests was comparable to the baseline model. This demonstrates the framework's effectiveness: although its training data (19,123 samples) was longer than the baseline data (11,851 samples) due to more clarification turns, leading to increased training time, the model still learned the core strategy of identifying high information-gain questions from these seemingly "inefficient" dialogues. **Filtering of Noise.** Nous-Noisy also performed nearly identically to the baseline model. This reveals a key property of our information-theoretic reward: it has a natural "immunity" to irrelevant information. Since noise cannot reduce the entropy of any attribute, its information gain reward is zero. Consequently, the training process automatically filters out the impact of noise, allowing the model to focus on learning genuinely effective question-answer patterns.

This study demonstrates our framework's high robustness to training data quality. Crucially, it also validates the robustness of our semantic parser, which successfully maps varied and imperfect responses back to the same underlying attributes, a key requirement for real-world application.

Table 5: Supplementary ablation study experimental results. Considering the significant time costs associated with data construction, model training, and drawing using VisPainter, the evaluation results of VisPainter are omitted in this experiment. Conduct model image generation evaluation experiments using nano-banana.

Method	Turns (↓)	Total IG (↑)	Human Judge(↑) Win	W/T(0.5)	W/T	GPT-5 Judge(↑) Win	W/T(0.5)	W/T
Nous-Template	20.3	**120.5**	0.29	**0.53**	**0.76**	**0.34**	**0.52**	0.69
Nous-Vague	22.1	117.1	**0.30**	0.50	0.69	0.29	0.50	**0.70**
Nous-Noisy	**19.7**	115.8	0.26	0.48	0.70	0.33	0.49	0.65

D Generalization Validation

Experimental Setting. To test whether our framework generalizes beyond scientific diagram generation, we evaluate it in collaborative novel writing. This domain differs substantially from diagram creation in both task structure and interaction dynamics, yet retains properties that make systematic study feasible. Novel writing is open-ended and creative, but it is also composed of recurring elements such as characters, settings, and events. These elements can be represented as structured state vectors, enabling the construction of a world model and the computation of per-turn information gain. At the same time, evaluation is relatively tractable: the quality of co-created narratives can be assessed through outline coverage and comparative judgments of readability and fidelity. These characteristics make collaborative novel writing another ideal testbed for examining the generality of our Socratic inquiry framework.

Data Preparation and Training. We collect novels from publicly available corpora. Since long-form narratives are often lengthy and would substantially increase the workload, we simplify the data by selecting representative chapters as test material, which are further rewritten through AI-assisted editing to avoid copyright concerns. In total, we obtain 120 processed samples, with 100 used for training and 20 for testing. From each sample, we extract structured elements such as characters, settings, conflicts, and resolutions to form state vectors and construct a world model as the prior. The data construction process follows the main text: the ground truth outline is provided to an Oracle, which answers model queries during simulation. Each question–answer pair is scored by information gain to create a preference dataset. Nous (OfG) is trained with offline GRPO, Nous (SFT) with supervised fine-tuning, and GPT baselines (zero-shot and few-shot: GPT-zero, GPT-fews) are included for evaluation.

Evaluation Metrics. For evaluation, we adopt two dimensions consistent with the main paper: interaction efficiency and output quality. Interaction efficiency is measured by dialogue turns and total information gain, reflecting whether a model can ask high quality questions within a limited number of turns. Output quality is assessed through outline coverage and subjective quality evaluation. Specifically, we compare the generated summaries of novel passages using both human and GPT judges in pairwise evaluations. These metrics provide a balanced view of how effectively the models gather information and how well they translate it into coherent creative output.

Results and Discussion. Novel writing represents a common and relatively structured domain, where LLMs already possess strong intrinsic capabilities. As shown in Table 6, this leads to notable efficiency for untrained models, which complete dialogues in fewer turns. However, Nous (OfG) achieves about 15% higher cumulative information gain compared to untrained baselines, confirming the benefit of entropy-based training. In terms of outline coverage, both OfG and SFT perform strongly, while GPT-few and GPT-zero show little distinction. For subjective evaluations by humans and GPT-5 judges, trained models consistently outperform baselines, though the margin is smaller than in our main domain. This may be due to the limited dataset size or the strong prior ability of LLMs in storytelling. Overall, the results validate that our framework retains effectiveness in a distinct creative domain, reinforcing its generalization capability and highlighting directions for future work in broader applications.

Table 6: Novel writing generalization experiment results. Dialogue efficiency and output quality are reported. All win-rate proportions are based on 80 pairwise judgments per model pair (20 prompts × 2 judges × 2 renderers).

Method	Turns (↓)	Total IG (↑)	Coverage (↑)	Human Judge(↑)			GPT-5 Judge(↑)		
				Win	**W/T(0.5)**	**W/T**	**Win**	**W/T(0.5)**	**W/T**
Nous (OfG)	14.2	**65.4**	**0.77**	**0.51**	**0.54**	**0.57**	**0.51**	**0.56**	**0.61**
Nous (SFT)	11.1	60.7	0.73	0.49	0.51	0.53	0.44	0.51	0.57
GPT-few	**10.4**	57.8	0.68	0.43	0.46	0.50	0.40	0.48	0.56
GPT-zero	13.7	55.2	0.67	0.46	0.49	0.51	0.37	0.45	0.53

E Detailed Introduction to the VisPainter Framework and In-depth Analysis of Experimental Results

The original intention of introducing the VisPainter framework for image drawing is to supplement the evaluation experiments. Because the comparative evaluation in Section 4.2.2 is based on subjective assessments by humans or AI, which can only subjectively measure the similarity between the Ground Truth and the generated images, making it difficult to quantify specific indicators. However, VisPainter can just provide an end-to-end process from generation to evaluation, and each indicator in its evaluation framework has a specific and quantifiable metric. This provides a perfect supplement to our experiments. But even if we only consider the experimental results in 4.2.2, they can still prove the effectiveness of the method proposed in this paper.

Introduction to the VisPainter Framework We adopt VisPainter as a baseline because it addresses a key limitation of diffusion based text to image models: instead of producing rasterized bitmaps, it generates fully editable vector diagrams. This property is crucial for scientific illustration, where accuracy, semantic clarity, and iterative refinement are essential.

VisPainter is a multi agent framework built on the Model Context Protocol (MCP). The workflow is organized into three roles. The Manager parses intent and coordinates tasks, the Designer drafts and refines layouts, and the Toolbox provides more than thirty MCP wrapped atomic drawing operations. In our experimental setup, GPT-4o is used as the Manager and gemini-2.5-pro as the Designer, while the Toolbox handles structured execution. These roles collaborate to translate natural language instructions into structured, editable diagrams through iterative refinement.

Furthermore, the module responsible for evaluation in the VisPainter framework is VisBench. VisBench is a benchmark for evaluating scientific schematics, providing 7 evaluation metrics across four dimensions, which are: accuracy, recall, design error rate, blank space rate, readability, alignment, and design steps. In our experimental evaluation, we only consider output quality, so the metric of design steps is not included. It is an additional reminder that the VisBench evaluation set currently contains 360 evaluation data entries, among which 180 are applicable to the T2I scenario and 180 to the TI2I scenario. The 100 test sets in our paper are selected from the T2I dataset among them. This integration makes VisPainter not only a generative system but also a research platform that supports rigorous and reproducible evaluation, making VisPainter a suitable and fair benchmark in our research. To the best of our knowledge, VisPainter was developed slightly earlier than or concurrently with our research work, and its open-source version will be released in the near future.

Analysis of Experimental Results In the results of Experiment 4.2.2, it can be observed that the design error rate score and alignment score show a trend inconsistent with other dimensions, and there is a situation where the simpler the output result, the higher the score. This is because the data counted in these two dimensions is related to the number of elements. The output elements of the untrained group are significantly fewer than those of the trained group. When the number of elements is significantly reduced, the error rate of the designer, namely gemini-2.5-pro, will increase slightly. That is to say, these dimensions test more the design ability of the designer, rather than the quality of the final integrated description output by the model. The shorter the final summary content output by the model and the fewer the elements, the fewer opportunities for mistakes in the drawing process, which in turn leads to the situation where the richer the output, the lower the score.

F Limitations and Future Work

Limitations **The Attribute Independence Assumption:** For computational tractability, we assume conditional independence between attributes. Although this is a reasonable and effective first-order approximation, many real-world tasks involve complex dependencies; for instance, a specific layout choice might constrain the types of available components. Our current model does not explicitly model these interactions, leaving this as a promising direction for future work.

Simulation-Based Evaluation: Although our automated simulation process can carry out large-scale and reproducible experiments, it cannot fully capture the complexity of human behavior. Real users may change their minds, express frustration, or continuously adjust their goals during the conversation. Therefore, it is necessary to conduct a comprehensive evaluation on human subjects in the next step to verify applicability in real-world scenarios.

Future Work The framework presented here has the potential to generalize to other structured domains, such as UI design, data visualization, or game creation. Beyond this broad applicability, two research directions are especially promising.

Learning the Task Space: Future agents could move beyond a fixed attribute set by inferring relevant attributes and their structure directly from interaction or large dialogue corpora. This would allow the framework to adapt dynamically to new tasks without manual specification.

Toward Mixed-Initiative Dialogue: Our current model is agent-led. A natural extension is to support mixed-initiative collaboration, where users proactively contribute information and the agent must decide whether to integrate it or pivot its strategy. This would yield more natural and adaptive interaction.

Together, these directions point toward making inquiry-driven collaboration more generalizable and human-like.

G Implementation Details

Training Environment and Hyperparameters All models were trained using a full-parameter fine-tuning approach on a high-performance computing cluster equipped with 8x NVIDIA H200 (141GB) GPUs. We utilized bfloat16 mixed-precision training to optimize for speed and memory efficiency. The key hyperparameters used for training each of the models are detailed in Table 7. We selected these parameters based on preliminary experiments to ensure stable and effective training for each respective method.

Table 7: Hyperparameters for SFT, DPO, OnG, and OfG.

Hyperparameter	**SFT**	**DPO**	**OnG**	**OfG**
Model & Data Configuration				
Base Model	Qwen3-8B			
Fine-tuning Method	Full-parameter			
Training Precision	bfloat16			
Max Sequence Length	4096			
Optimization				
Optimizer	AdamW			
Learning Rate (lr)	1e-6	1e-6	1e-6	**1e-6**
LR Scheduler Type	Cosine			
Warmup Steps	50	50	50	**50**
Epochs	5	5	5	**5**
Batch Size (per device)	1	1	1	**1**
Gradient Accum. Steps	2	2	2	**2**
Weight Decay	0.01	0.01	0.01	**0.01**
Regularization & RL-specific				
KL Coefficient (β)	N/A	0.1	0.01	**0.01**
PPO Clip Epsilon (ϵ)	N/A	N/A	0.2	**0.2**

H Dataset Details

Our dataset was constructed from a corpus of scientific papers sourced from arXiv and PubMed Central (PMC), covering a wide range of disciplines to ensure diversity. The primary arXiv categories included Computer Science (43.1%), Physics (22.7%), Quantitative Biology (14.8%), Electrical Engineering (11.5%), and others such as Economics and Statistics (7.9%). All source materials were confirmed to be under open-access licenses (e.g., Creative Commons, arXiv.org non-exclusive license) that permit reuse for research. The initial pool of approximately 1 million figures was refined through a multi-stage pipeline: an initial filtering with CLIP to remove data plots, followed by a fine-grained selection of schematic diagrams using Qwen-2.5-VL-72B. A final manual verification by three domain experts ensured the relevance and quality of each diagram, resulting in a curated set of 1,100 figures. Of these, 1,000 were used for training and 100 were held out for testing. To ensure the reproducibility of our experiments involving proprietary models, all API calls for data generation and evaluation were made using model versions available after 4-14-2025.

I prompts

All prompt templates, data construction, model training, and result evaluation codes are included in the attachments submitted along with the article. Here we provide the Socratic prompting templates used for our zero-shot and few-shot baselines. The core idea is to encourage the assistant to proactively ask clarifying questions before finalizing the diagram specification:

I.1 Zero-Shot Socratic Prompting

```
"You are an assistant that helps design scientific diagrams.
Do not produce the diagram immediately. Instead, follow these steps:
1. Ask the user a clarifying question about the diagram (e.g., type,
    layout, number of components, connections, or style).
2. Continue asking such clarifying questions until enough information has
     been gathered to produce a complete diagram specification.
3. Only after clarification is complete, summarize the final diagram
    specification in a structured format (JSON).
Remember:
- Ask focused, concrete questions (one per turn).
- Avoid vague or open-ended questions.
- The final specification must be complete and self-contained."
```

I.2 FEW-SHOT SOCRATIC PROMPTING

```
"You are an assistant that helps design scientific diagrams.
Do not produce the diagram immediately. Instead, follow these steps:
1. Ask the user a clarifying question about the diagram (e.g., type,
    layout, number of components, connections, or style).
2. Continue asking such clarifying questions until enough information has
     been gathered to produce a complete diagram specification.
3. Only after clarification is complete, summarize the final diagram
    specification in a structured format (JSON).
Remember:
- Ask focused, concrete questions (one per turn).
- Avoid vague or open-ended questions.
- The final specification must be complete and self-contained.
Here are two example questions for your reference:Example 1: Example 2:"
```

J DATA AND ILLUSTRATION EXAMPLE

Data Example Example of offline GRPO training data based on information entropy:


```
"  {
    "prompt": "User: I want to create a scientific diagram.\nAssistant:
    What specific element or component should serve as the target
    identifier (connection.to_id) for the connections ...",
    "responses": [
      "What is the primary scientific concept this diagram aims to
    illustrate?",
      "What is the primary scientific concept or model you intend to
    illustrate with these components and connections?",
      "What is the fundamental scientific concept or process that your
    diagram is intended to illustrate?",
      "What primary scientific concept or system does this diagram aim to
     illustrate?",
      "What is the primary scientific concept or system that this diagram
     is intended to illustrate?",
      "What overall visual structure do you envision for your diagram?",
      "Which overall diagram type do you envision such as a flowchart
    ...?",
      "What specific element or component should ...?"
    ],
    "reward": [
      16.83,
      13.91,
      10.56,
      8.21,
      6.78,
      4.09,
      1.32,
      0.0
    ]
  },"
```


Example of offline GRPO training data based on slot-counting:


```
"    {
    "prompt": "User: I want to create a scientific diagram.",
    "responses": [
      "What is the primary scientific concept this diagram aims to
    illustrate?",
      "What is the primary scientific concept or topic that this diagram
    is intended to illustrate?",
      "What is the primary scientific concept or topic that this diagram
    aims to illustrate?",
```

```
    "What are the main components of your scientific diagram, and what
  unique identifiers (component.id) will you assign to each?",
    "What specific components (component.id) do you envision in your
  diagram, and what identifiers or labels should be assigned to each to
   clarify their roles?",
    "Which key components (nodes) do you envision for your diagram, and
   how would you uniquely identify each (i.e., what are their
  respective component IDs)?",
    "Can you identify the distinct components for your diagram by
  assigning specific IDs or names, and briefly describe the role of
  each?",
    "What are the main components (component.id) you envision including
   in your scientific diagram, and what specific role does each play in
   illustrating the concept?"
  ],
  "reward": [
    4.0,
    2.0,
    2.0,
    1.0,
    1.0,
    1.0,
    1.0,
    1.0
  ]
},"
```


Objective metrics This section presents drawing examples generated using the VisPainter framework, as shown in Figure 5.

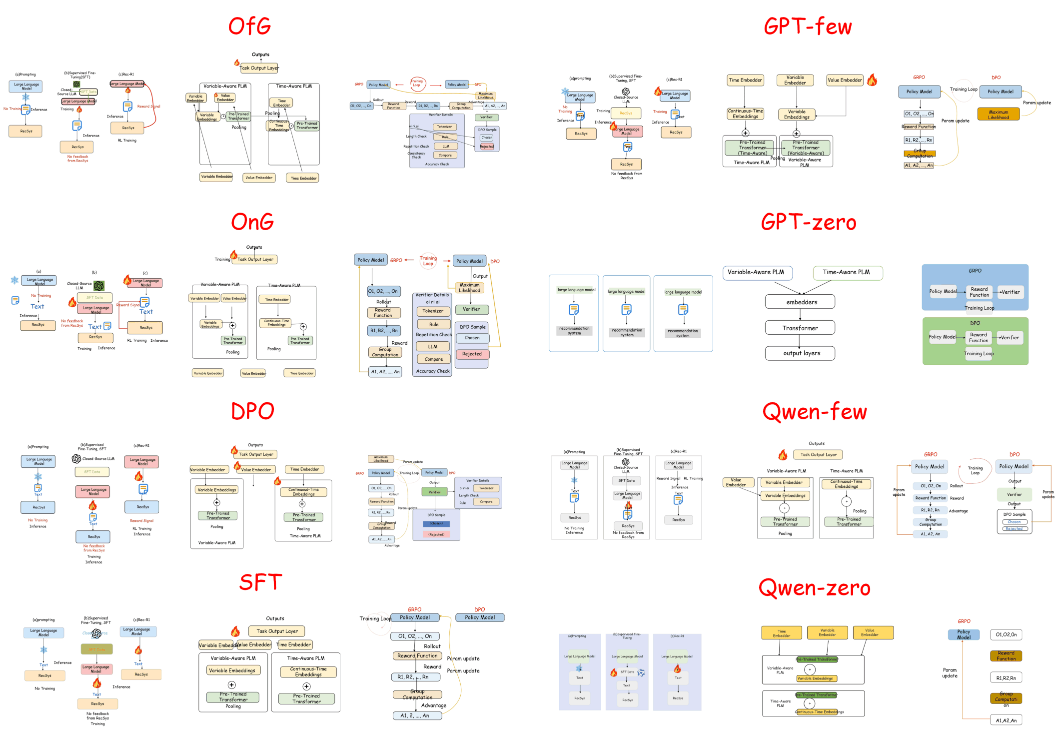

Figure 5: This section presents drawing examples generated using the VisPainter framework

Subjective comparison This section presents drawing examples generated by two models (4o-image-1 and nano-banana), as shown in Figure 6.

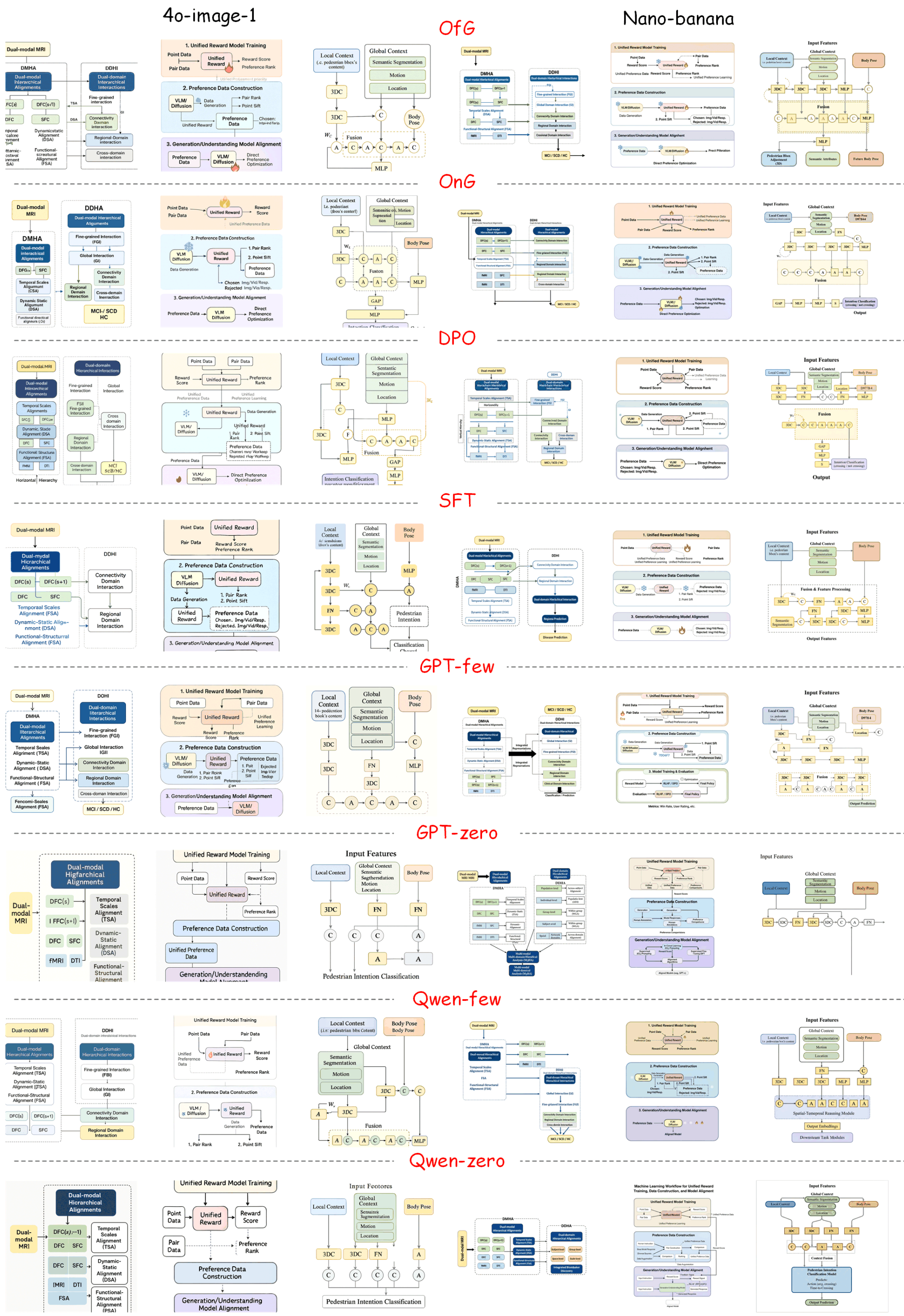

Figure 6: Partial Examples of Model-Generated Images