Designed to Spread: A Generative Approach to Enhance Information Diffusion

Ziqing Qian[*1,2], **Jiaying Lei**[*1,3], **Shengqi Dang**[1,2,3], **Nan Cao**[1,2,3†]

[1]Intelligent Big Data Visualization Lab, Tongji University, Shanghai, China
[2]Shanghai Research Institute for Intelligent Autonomous System, Tongji University, Shanghai, China
[3]Shanghai Innovation Institute, Shanghai, China
2411920@tongji.edu.cn, jiaying.lei@outlook.com, dangsq123@tongji.edu.cn, nan.cao@gmail.com

Abstract

Social media has fundamentally transformed how people access information and form social connections, with content expression playing a critical role in driving information diffusion. While prior research has focused largely on network structures and tipping point identification, it provides limited tools for automatically generating content tailored for virality within a specific audience. To fill this gap, we propose the novel task of *Diffusion-Oriented Content Generation* (DOCG) and introduce an information enhancement algorithm for generating content optimized for diffusion. Our method includes an influence indicator that enables content-level diffusion assessment without requiring access to network topology, and an information editor that employs reinforcement learning to explore interpretable editing strategies. The editor leverages generative models to produce semantically faithful, audience-aware textual or visual content. Experiments on real-world social media datasets and user study demonstrate that our approach significantly improves diffusion effectiveness while preserving the core semantics of the original content.

Code — https://github.com/idvxlab/designed-to-spread

Introduction

Social media has reshaped how people communicate, access information, and form connections. Platforms such as X, Facebook, and TikTok have become powerful channels for news dissemination, public discourse, and cultural influence. Information spreads via diffusion-like processes along user networks, motivating extensive research on diffusion mechanisms (Bakshy et al. 2012; Guille et al. 2013; Weng, Menczer, and Ahn 2013).

Extensive research has examined information diffusion in social networks, including modeling diffusion dynamics (Granovetter 1978; Pastor-Satorras and Vespignani 2001; Kempe, Kleinberg, and Tardos 2003), identifying influential seed users (Chen, Wang, and Yang 2009), and limiting spread via edge removal (Tong et al. 2012). These approaches typically depend on detailed network structure, which is often unavailable or incomplete in real-world settings. Meanwhile, how content is crafted also significantly affects its diffusion, especially for specific audiences. Yet, content-level optimization remains underexplored. This highlights the need for automated methods to generate audience-aware, diffusion-oriented content without relying on explicit network information.

The rapid development of AIGC techniques offers promising potential to fulfill the above needs. While these methods excel at generating high-quality text and image content (Chen et al. 2023; Podell et al. 2023), optimizing content specifically for maximizing information diffusion remains challenging due to three key issues: (1) diffusion impact is difficult to measure at the content level, especially when network topology is unavailable or incomplete; (2) preserving the original message's intent while generating user-preferred representations is non-trivial, as misalignment may cause misunderstanding or even unintentional misinformation; and (3) designing a unified framework that integrates diffusion feedback into multimodal content generation remains an open challenge.

To address the above challenges, we introduce the task of *Diffusion-Oriented Content Generation* (DOCG), which aims to generate a semantically faithful variant (text or image) of a given message on a specific topic, optimized to maximize its diffusion impact among a target audience group. We propose an information enhancement framework with two key components: an *influence indicator* and an *information editor*. The influence indicator estimates content-level diffusion potential within the target audience. The information editor formulates content generation as a reinforcement learning (RL) problem, guiding a generative model to revise content through interpretable, modality-specific editing actions (e.g., amplifying emotional tone). The RL objective is to maximize predicted influence while preserving the original message's semantics. Our main contributions are summarized as follows:

- We propose a new task, *Diffusion-Oriented Content Generation* (DOCG), which aims to generate audience-aware content to maximize diffusion influence.
- We propose a reinforcement learning framework that reshapes input messages to enhance diffusion while preserving their original intent. It leverages an influence indicator to estimate content-level impact for target audiences without relying on network topology, using this

[*]These authors contributed equally.
[†]Corresponding author.

feedback to guide iterative multimodal revisions.

- We demonstrate the versatility of our method across both textual and visual modalities, achieving significant improvements on real-world social media dataset.

Related Work

In this section, we review related work in three areas most relevant to our study: social media analysis, information diffusion models, and content editing and generation.

Social Media Analysis

Social media analysis explores the complex interplay of factors that shape how information is created, interpreted, and propagated online. Early studies primarily adopted descriptive approaches to reveal observable trends (Cheng et al. 2014; De Francisci Morales, Monti, and Starnini 2021; Etta et al. 2023; Flamino et al. 2023), drawing insights from three main perspectives.

The first focuses on content virality. For example, Vosoughi et al. (Vosoughi, Roy, and Aral 2018) found that false information spreads more rapidly than truthful content on X, largely due to its novelty and emotional appeal. The second examines user engagement behaviors, including posting frequency (Benevenuto et al. 2009; Spasojevic et al. 2015), reaction latency (Hodas and Lerman 2014), and cross-platform activity (Xu et al. 2014; Iamnitchi et al. 2023; Alipour et al. 2024). Especially, Shahbaznezhad et al. (Shahbaznezhad, Dolan, and Rashidirad 2021) further demonstrated that content format and platform-specific algorithms significantly affect user responsiveness, with emotionally charged posts eliciting faster reactions. The third investigates the structural properties of diffusion cascades. Notarmuzi et al. (Notarmuzi et al. 2022) identified universal propagation patterns in early retweet cascades, highlighting the role of network topology and user influence in determining final reach and virality.

While these studies offer valuable insights into diffusion dynamics, they remain largely observational and often overlook the causal impact of content transformation on user behavior. This gap motivates our generative approach to enhancing information diffusion through content design.

Information Diffusion Model

Research on information diffusion models seeks to model the content propagation through social networks. Early approaches were inspired by epidemiological processes, such as the Independent Cascade (IC) and Linear Threshold (LT) models; however, these models assume static networks and fixed influence probabilities among different people (Goldenberg, Libai, and Muller 2001; Kempe, Kleinberg, and Tardos 2003).

To overcome these simplifications, subsequent work has incorporated temporal dynamics, deep learning, and evolving network structures. For example, NetRate (Gomez-Rodriguez, Leskovec, and Krause 2012) applies survival analysis to infer time-dependent transmission functions from cascade data. Inf-VAE (Sankar et al. 2020) models latent diffusion representations via variational autoencoders, while DeepDiffuse (Islam et al. 2018) leverages deep neural networks to predict both the participants and timing of information cascades. More recently, Meng et al. (Meng et al. 2025) propose a data-driven model that captures the dynamics of social reinforcement and decay in online diffusion, offering a concise formulation for large-scale spread.

Despite these advancements, existing models still exhibit critical limitations: a strong reliance on explicit network structures, sensitivity to seed selection, and a lack of content awareness in modeling (Jiang, Ren, and Ferrara 2023). These challenges underscore the need for a content-based diffusion influence indicator that functions independently of network topology.

Content Editing and Generation

Content editing and generation synthesize new content aligned with a target distribution or specific objectives such as style transfer. Generation creates content from scratch, while editing modifies existing references to achieve desired transformations, preserving coherence and semantics.

Early work on content generation used variational autoencoders (VAEs) (Kingma, Welling et al. 2013) and generative adversarial networks (GANs) (Goodfellow et al. 2020) to learn latent representations for attribute-controlled editing (Radford, Metz, and Chintala 2016; Dumoulin et al. 2017). With large-scale pretrained models, generation has become more fluent and controllable. Large language models (LLMs) such as GPT-3 (Brown et al. 2020) and T5 (Raffel et al. 2020) enable fine-grained text editing via prompts. In vision, diffusion models including Stable Diffusion (Rombach et al. 2022), SDXL (Podell et al. 2023), and PixArt-α (Chen et al. 2023) support high-quality image generation and editing. Multimodal systems such as BLIP-Diffusion (Li, Li, and Hoi 2023) and MIGE (Tian et al. 2025) further unify the editing across modalities.

However, most methods focus on general-purpose generation. Although recent work explores engagement-driven generation for social networks (Coppolillo et al. 2025), audience-specific preferences remain overlooked. Leveraging the adaptability of LLMs and diffusion models, we explore audience-aware content generation to better support information diffusion.

Diffusion-Oriented Content Generation

In this paper, we introduce a new task, *Diffusion-Oriented Content Generation* (DOCG), which aims to generate a semantically faithful variant (text or image) of a given message c, optimized to maximize its diffusion impact among a group of target audiences U. We also propose a novel modality-free algorithm framework based on reinforcement learning to accomplish this task. The core idea of our framework is to evaluate the diffusion impact and select strategies to edit the original message iteratively based on pretrained generative models such as GPT-4 (for text) or Pixart-α (Chen et al. 2023) (for image). By repeating this evaluation–editing loop, the framework progressively enhances the expected diffusion impact, ultimately producing an optimized variant tailored for the group of target audiences.

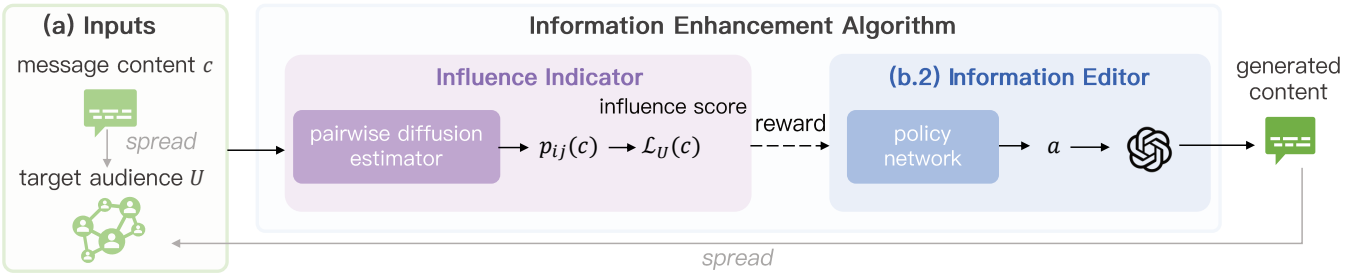

Figure 1: Overview of our proposed information enhancement framework, consisting of two main components: the influence indicator, and the information editor.

As shown in Figure 1, our framework accepts an initial message c and a group of target audiences U as the inputs, and comprises two components: (b.1) the *influence indicator* and (b.2) the *information editor*. The influence indicator computes an overall influence score $\mathcal{L}_U(c)$ to measure the diffusion impact. The information editor treats $\mathcal{L}_U(c)$ as the reward signal for a policy network that samples an editing action a. This action is executed by a generative agent (e.g., a large language or text-to-image model) to produce a revised message. In the following sections, we describe the modeling and implementation details of each component.

Influence Indicator

The influence indicator produces an influence score to quantify the diffusion impact of the message content c within the group of target audiences U. Formally, the influence score is defined as:

$$\mathcal{L}_U(c) = \max_i \Big(\frac{1}{|U|} \sum_{u_j \in U} p_{ij}(c) \Big) \tag{1}$$

where $p_{ij}(c)$ denotes the probability that user $u_j \in U$ will spread (e.g., retweet or reply to) content c after being exposed to it from user $u_i \in U$. This formulation captures the diffusion potential of the most influential initiator. In the following, we detail the pairwise diffusion estimator for computing $p_{ij}(c)$, its training data, and its loss function.

Pairwise Diffusion Estimator. The pairwise diffusion estimator is implemented as a neural network that predicts the probability $p_{ij}(c)$. The model operates in three stages: feature extraction, projection, and interaction. First, we extract the features f_i, f_j, and $f(c)$ corresponding to users u_i, u_j, and content c, respectively. User features are computed as the average of features from previously created or shared content, while content features are obtained using a pre-trained CLIP encoder (Radford et al. 2021), which aligns the text and image features in a shared space. To handle long-text inputs, we adopt Long-CLIP (Zhang et al. 2024), a variant of CLIP designed to overcome token length limitations. Then, all input features are projected into a latent space via a shared Multilayer Perceptron (MLP), denoted as M_1. The combined representation h is obtained by concatenating the projected vectors:

$$h = M_1(f_i) \oplus M_1(f_j) \oplus M_1(f(c)) \tag{2}$$

where $\oplus$ denotes vector concatenation. Finally, h is passed through another MLP, M_2, which captures nonlinear interactions among the user and content features. A sigmoid function $\sigma(\cdot)$ is applied to obtain the final probability:

$$p_{ij}(c) = \sigma\big(M_2(h)\big) \tag{3}$$

Training Data. We construct the training data not rely on explicit social network topology by assuming a fully connected graph over the audiences of each content. We formulate each training sample as a quadruple (u_i, u_j, c, y_{ijc}), where u_i and u_j are user pairs, c is the content, and $y_{ijc} \in \{0, 1\}$ is the label showcasing the interaction behavior. For a given content c, let u_o denote the original poster and U_c the set of users who interacted with u_o (e.g., via retweet or reply). A sample is labeled as positive ($y_{ijc} = 1$) if either (1) $u_i = u_o$ and $u_j \in U_c$, or (2) both $u_i, u_j \in U_c$. All other user pairs are labeled as negative ($y_{ijc} = 0$), indicating no observed interaction related to c between u_i and u_j. To mitigate class imbalance, we perform negative sampling by randomly dropping a subset of training samples with $y_{ijc} = 0$, ensuring that the number of negative samples approximately matches the number of positive ones for each content.

Loss Function. To train the pairwise diffusion estimator, we minimize a binary cross-entropy loss, defined as:

$$\mathcal{L}_{CE} = \frac{1}{M} \sum_{i,j,c} [y_{ijc} \log p_{ij}(c) + (1 - y_{ijc}) \log (1 - p_{ij}(c))] \tag{4}$$

where M is the total number of training instances, $p_{ij}(c)$ is the output of the predicted diffusion estimator, and $y_{ijc} \in \{0, 1\}$ is the groundtruth label. The first term penalizes the model when it assigns a low probability to observed diffusion events, encouraging high confidence in positive cases. The second term penalizes false positives, pushing the model to assign low scores when diffusion does not occur.

Information Editor

Figure 2 illustrates the algorithm workflow of the information editor introduced in our framework. It rewrites the text or image content c to improve its diffusion impact. The content c, along with its CLIP features $f(c)$, is treated as the *state* s. Based on the state, the policy network π_θ, parameterized by θ, produces a conditional distribution $\pi_\theta(a|s)$ over editing actions and samples an interpretable *action* vector a. This action is then passed to an *agent* (i.e., GPT-4 for text

	Dimension	Description
Text	*Social Currency*	whether the text enhances the sharer's profile (e.g., appearing good, intelligent, funny).
	Triggers	whether the text is strongly connected to real-life scenarios, promoting frequent recall.
	Emotion	whether the text evokes high-arousal emotions (e.g., excitement, surprise, delight).
	Public	whether the text format is easily shareable and visible.
	Practical Value	whether the text offers concrete guidelines or useful information.
	Stories	whether the text includes a narrative framework or memorable elements.
Image	*Colorfulness*	whether the image is rich in color and has visual impact.
	Human Scene	whether the image includes scenes with people to enhance emotional resonance.
	Emotion	whether the image can evoke strong emotional responses (such as joy, shock, or being moved).
	Professional	whether the image has the quality of professional photography.
	Brightness	whether the image is bright and attention-grabbing.
	Clarity	whether the image is clear and its details are prominent.
	Visual Balance	whether the visual elements in the image are evenly distributed.
	Focus of the Picture	whether the image focuses on the subject, avoiding visual dispersion.

Table 1: Action space for text and image content generation.

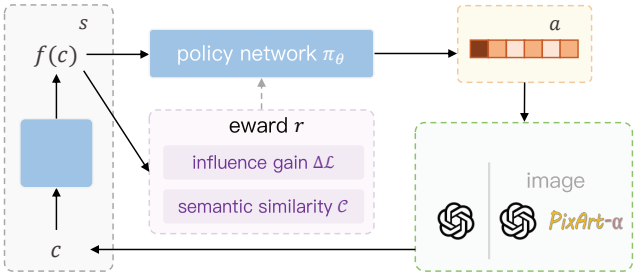

Figure 2: Our information editor for editing message content c. The policy network generates editing actions to iteratively revise the content, guided by the reward signal.

revision or GPT-4 combined with PixArt-α for image generation), which produces the revised content representing the next state. To train the policy network and guide the rewriting process toward maximizing diffusion impact, a *reward* is computed by combining two parts: (i) influence gain $\Delta\mathcal{L}$, quantifying the improvement of the influence score as estimated by the influence indicator, and (ii) semantic similarity $\mathcal{C}$, measuring semantic consistency of the rewritten content and the original message. We describe each of these algorithm details in the rest of the section.

Action. We define an interpretable action vector a, where each dimension $a_i \in [-1, 1]$ (as detailed in Table 1) corresponds to a specific content feature to be modified. The value of a_i indicates the degree and direction of modification: $a_i \approx +1$ enhances the i-th feature, $a_i \approx 0$ leaves it unchanged, and $a_i \approx -1$ suppresses it. For text content, we adopt the STEPPS framework (Pressgrove, McKeever, and Jang 2018), which includes six dimensions—*Social Currency*, *Triggers*, *Emotion*, *Public*, *Practical Value*, and *Stories*—each empirically linked to increased virality. For image content, the action space comprises eight perceptual dimensions related to visual engagement, such as *Colorfulness*, *Brightness*, and *Visual Balance*. These dimensions are informed by Li and Xie (Li and Xie 2020), who identified image characteristics (e.g., color variation, emotional expression, and photographic quality) as key drivers of user engagement on social media.

Agent. The agent leverages prompt engineering to rewrite content using large language models, such as GPT-4 in our implementation (implementation details are available in the Appendix). Briefly, we embed the editing *action* into the prompt, guiding the model to perform controlled rewriting. For text content, we construct a dynamic prompt that encodes the action vector values (ranging from -100% to 100%) across six STEPPS dimensions. Each dimension is associated with a natural language instruction, annotated with the action vector value. For image content, we adopt a similar strategy to construct the text-to-image prompt, guided by eight predefined visual dimensions. The resulting prompt is then input into the image generation model (e.g., PixArt-α in our implementation), enabling controlled manipulation of visual features that affect diffusion.

Reward. The reward quantifies the quality of a revision by jointly considering diffusion improvement and semantic fidelity. It is defined as:

$$r = \begin{cases} \sqrt{\Delta\mathcal{L} \cdot \mathcal{C}} & \Delta\mathcal{L} \geq 0 \\ -\sqrt{-\Delta\mathcal{L} \cdot (1 - \mathcal{C})} & \Delta\mathcal{L} < 0 \end{cases} \tag{5}$$

where the influence gain $\Delta\mathcal{L}$ and the semantic similarity $\mathcal{C}$ are defined as:

$$\Delta\mathcal{L} = \mathcal{L}_U(c) - \mathcal{L}_U(c^{\text{ori}}) \tag{6}$$

$$\mathcal{C} = \cos(f(c), f(c^{\text{ori}})) \tag{7}$$

where c^{ori} is the original input message content, cos denotes the cosine similarity. This formulation rewards the edits that boost diffusion while preserving semantic fidelity, and penalizes changes that are either semantically inconsistent or reduce diffusion potential.

Policy Network. The policy network defines a conditional distribution $\pi_\theta(a|s)$ over editing actions a, given the current state s, represented by the content feature vector $f(c)$. To generate a continuous action vector $a \in [-1, 1]$, we use a reparameterization strategy:

$$a = \tanh\left(\mu_\theta(c) + \sigma_\theta(c) \cdot \epsilon\right) \tag{8}$$

Dataset	Seed Tweets	Users	All Tweets	Time Period
Olympics	4,375	10,097	5,812,520	2024.8-2024.11
Movie	6,625	14,552	6,181,779	2022.4-2024.11

Table 2: Statistics of datasets.

Dataset	Precision	Recall	F1-score	AUC
Olympics	0.8187	0.8122	0.8113	0.9042
Movie	0.8360	0.8286	0.8277	0.9186

Table 3: Performance of the pairwise diffusion estimator.

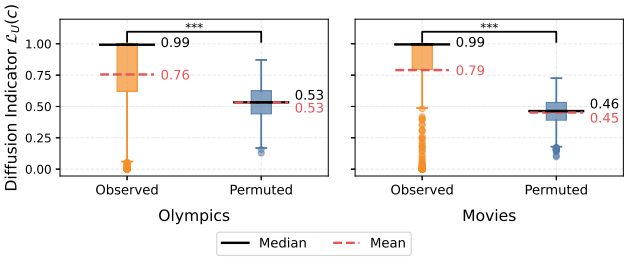

Figure 3: Distribution of influence scores for messages on *observed* (orange) and *permuted* (blue) combinations. *** denotes statistical significance at $p < 0.001$.

where $\epsilon \sim \mathcal{N}(0, I)$ is standard Gaussian noise that introduces controlled stochasticity to encourage exploration. The hyperbolic tangent function $\tanh(\cdot)$ bounds the output within $[-1, 1]$, promoting numerical stability and interpretability. The mean $\mu_\theta(c)$ and standard deviation $\sigma_\theta(c)$ are predicted by a two-layer MLP applied to $f(c)$, enabling the policy to flexibly capture uncertainty in the action space.

The objective is to learn an optimal policy that maximizes the expected reward:

$$\pi^*(a|s) = \arg\max_{\pi_\theta} \sum_a \pi_\theta(a|s) \cdot Q(s, a) \tag{9}$$

where $Q(s, a)$ is the expected reward for applying action a to state s, and π_θ is the policy parameterized by θ. We optimize the policy network using the policy gradient method (Sutton et al. 1999), which updates parameters to maximize the expected cumulative reward:

$$\theta \leftarrow \theta + lr \cdot \frac{1}{B} \sum_{i=1}^{B} \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_i^{(t)}|s_i^{(t)}) \cdot J_{i,t} \tag{10}$$

where i indexes the sample in the batch, $a_i^{(t)}$ is the action taken at step t, $s_i^{(t)}$ is the state at step t, ∇_θ is the gradient with respect to policy parameters θ, B is the batch size, γ is the discount factor, lr is the learning rate, and $J_{i,t}$ is the discounted return from step t, defined as:

$$J_{i,t} = \sum_{k=t}^{T} \gamma^{k-t} r_{i,k} \tag{11}$$

in which $r_{i,k}$ is the reward at step k computed by Eq. (5).

Evaluation

We conducted comprehensive experiments to evaluate the effectiveness of our proposed framework. In this section, we begin by describing the dataset details. We then separately evaluate the influence indicator and the information editor. Finally, we present results from a user study designed to assess whether the generated content aligns with human preferences in the real-world information diffusion scenarios.

Datasets

We constructed two X (Twitter) datasets focused on discussions about recent movies and the Paris 2024 Summer Olympics (Table 2) using a three-step process. First, we collected tweets containing topic-relevant keywords, referred to as *seed tweets*. Second, for each seed tweet, we defined its audience group as the original poster and users who directly interacted with it via retweets or replies. Third, we merged all audience groups to form a unified topic-specific audience group, and collected their tweets within a defined time window to build the final dataset. In total, the datasets comprise over 10,000 users and 11 million tweets.

Indicator Estimation

In this section, we present the implementation and experimental setup of the influence indicator, followed by an evaluation of its effectiveness from two perspectives: (1) the predictive capacity of the pairwise diffusion estimator $p_{ij}(c)$, and (2) the ability of the influence score $\mathcal{L}_U(c)$ to capture variations in diffusion impact across different audience groups. We also provide qualitative examples of content with high and low influence scores to demonstrate the practical utility of the proposed indicator.

Implementation & Settings. We trained the pairwise diffusion estimator using the Adam optimizer with a learning rate of 1×10^{-4}. Input features were standardized by removing the mean and scaling to unit variance along each dimension. The dataset was split into training and test sets in a 4:1 ratio, with no overlap in users or messages between the two sets.

Results. As shown in Table 3, the pairwise diffusion estimator demonstrates strong predictive performance, achieving AUC scores of 0.9042 on Olympics and 0.9186 on Movie, indicating its effectiveness in distinguishing diffusion from non-diffusion behaviors. To evaluate the influence score $\mathcal{L}_U(c)$, we compare its values on *observed* message–audience combinations (c_i, U_i) against *permuted* combinations (c_i, U_j) with $i \neq j$, where U_i is the actual audience group for message c_i and U_j is unrelated. Figure 3 shows that observed combinations yield significantly higher scores (mean: 0.76 vs. 0.53 for Olympics; 0.79 vs. 0.45 for Movie), with minimal overlap between distributions. A Mann–Whitney U test confirms these differences are highly significant ($p < 0.001$), demonstrating that $\mathcal{L}_U(c)$ effectively captures audience-specific diffusion potential.

Visualization. Figure 4 presents representative messages from the Olympics dataset with low and high influence

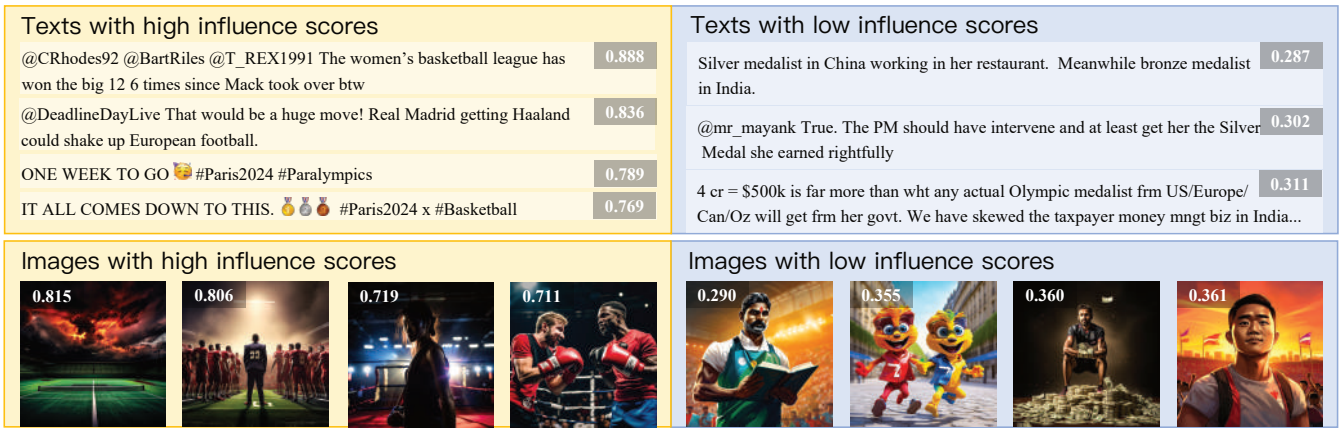

Figure 4: Representative text and image examples across different levels of influence scores $\mathcal{L}_U(c)$.

scores. Low-scoring examples typically lack topical relevance or suffer from weak linguistic or visual quality. In contrast, high-scoring messages exhibit strong contextual alignment, coherent composition, and features (e.g., emotion) that are more likely to engage the target audience.

Editor Estimation

In this section, we firstly describe the implementation details and experimental setup of information editor. We then outline the baselines and evaluation metrics, and present both quantitative and qualitative comparisons with baselines. Finally, we conduct an ablation study on the reward function.

Implementation & Settings. We trained the policy network for 350 episodes with a trajectory length of 3, using the Adam optimizer with a learning rate of 1×10^{-4}. For image content generation, each input message was first converted into a text-to-image prompt by GPT-4 and then processed by PixArt-α to generate the initial image. For evaluation, we sampled 200 messages from the Olympics test set and computed their influence scores. The 40 messages with the lowest scores were used as the test set, while the remaining 160 messages formed the training set.

Baselines. We compare our framework with the following baselines (detailed implementation could be found in the appendix): (1) *Large Language Model (LLM)*: directly prompts a pretrained language model to rewrite the input message, without any audience-specific guidance; (2) *In-Context Learning (IC-L)*: for each audience group, retrieves messages with the highest and lowest observed influence scores as positive and negative exemplars, respectively, and uses them as few-shot prompts to guide LLM-based rewriting; (3) *Greedy Search (Greedy)*: as a non-learning variant of our framework, uniformly samples editing actions and selects the revised message with the highest reward at each step, without learning a policy network.

Metrics. We assess our information editor using two metrics: (1) *Diffusion Gain*, which measures the relative increase of influence scores after rewriting, and (2) *Consistency*, which quantifies semantic preservation between original and edited content via the cosine similarity of their features. The details can be found in the appendix.

		Diffusion Gain↑	Consistency ↑
Text	**Ours**	**20.69%**	0.9510
	LLM	14.03%	**0.9590**
	IC-L	13.43%	0.9509
	Greedy	17.93%	0.9470
Image	**Ours**	**11.20%**	0.8730
	LLM	1.36%	0.8743
	IC-L	3.05%	**0.8748**
	Greedy	7.93%	0.8693

Table 4: Quantitative comparison with baselines.

Quantitative Comparison. Table 4 presents the performance on both text and image generation tasks. In the text task, our method achieves the highest Diffusion Gain at 20.69%, significantly outperforming other baselines. Additionally, our method attains a high Consistency score (0.9510), indicating effective semantic preservation. Although LLM achieves slightly higher Consistency (0.9590), its diffusion improvement is considerably weaker. For the image task, our method again leads with a Diffusion Gain of 11.20%, surpassing other baselines. In terms of Consistency, IC-L (0.8748) and LLM (0.8743) slightly exceed our score (0.8730). However, the difference is minimal, confirming that our edits maintain semantic integrity while delivering significantly better diffusion performance.

Qualitative Comparison. Figure 5 illustrates representative generation results for both text and image content. In the text examples, our framework transforms generic statements into vivid, action-oriented language that heightens emotional engagement and topical relevance while preserving the original intent. In the image examples, our approach enhances visual appeal by emphasizing human subjects, resulting in more compelling and audience-aware imagery.

Ablation Study. We assess the reward function by comparing the variant optimizing only the influence gain component $\Delta\mathcal{L}$. Table 5 shows the results. When trained with only the influence gain, the model boosts diffusion but sacrifices se-

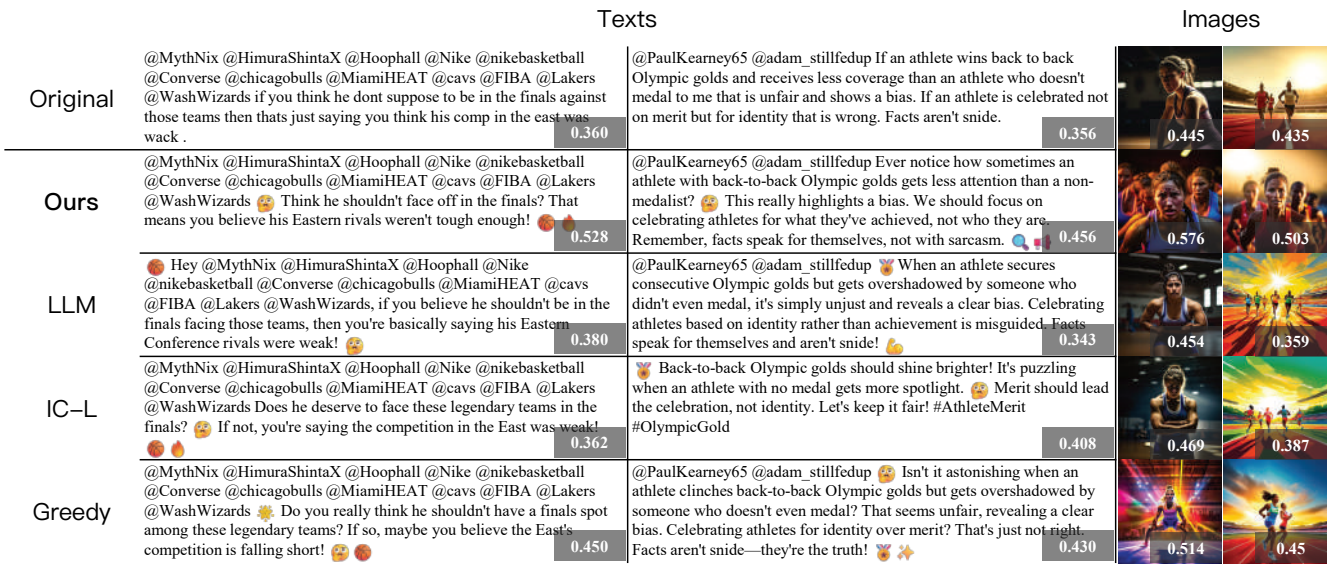

Figure 5: Qualitative comparison with baselines. Each piece of content is annotated with its corresponding influence score.

		Diffusion Gain↑	Consistency ↑
Text	$\sqrt{\Delta\mathcal{L}\cdot\mathcal{C}}$	20.69%	**0.9510**
	$\Delta\mathcal{L}$	**21.98%**	0.9471
Image	$\sqrt{\Delta\mathcal{L}\cdot\mathcal{C}}$	11.20%	**0.8730**
	$\Delta\mathcal{L}$	**12.43%**	0.8707

Table 5: Results of ablation study.

mantic accuracy. This highlights the importance of incorporating semantic similarity, as it ensures that the content remains meaningful while still achieving improved diffusion.

User Study

We conducted a user study to evaluate whether the generated content increases users' likelihood of sharing information in realistic settings.

Procedure & Metric. We conducted a user study on the Prolific platform (Prolific 2025), recruiting native English speakers with informed consent. Participants were screened based on their X activity and recent engagement with the target topic. A total of 100 qualified users were selected (76 male, 24 female; ages: 18–25 (11%), 26–50 (68%), 51–60 (21%)). Each participant evaluated 100 tweet pairs and 100 image pairs, each consisting of an original and a revised version of either text or visual content. For each pair, participants selected the version they were more likely to share (via retweet or reply). To mitigate ordering bias, all pairs were presented in randomized order.

We evaluated performance using the ***Retweet Preference Rate* (RPR)**—the average proportion of times the rewritten version was preferred by our participants.

Results & Analysis. The results indicate that the generated content significantly increased participants' willingness to retweet. The overall RPR was **62.36%**(**64.00%** for text and **60.72%** for image), suggesting that enhancements in both modalities contribute to improved diffusion impact. A paired-sample t-test confirmed that the overall preference for the rewritten content is statistically significant ($p < 0.001$). We further analyzed participants' stated reasons for their choices. Common reasons included stronger emotional resonance and alignment with personal experiences.

Discussion and Conclusion

In this work, we introduce a novel task, *Diffusion-Oriented Content Generation*, and propose a reinforcement learning (RL)–based framework that rewrites messages to enhance diffusion among targeted audiences. Central to our approach is an influence indicator that estimates the diffusion potential of content for a given audience, without relying on explicit network topology. Based on this indicator, we develop an information editor that optimizes message rewrites for maximal diffusion. We model the editor as a finite-horizon MDP and adopt a stationary policy for practicality, which remains effective even though optimal policies are generally non-stationary. Experiments on real-world datasets show that our indicator accurately predicts diffusion outcomes, and that the RL-based rewriting strategy yields significant diffusion gains while maintaining semantic fidelity.

Despite promising performance, our method still has limitations. Our current iterative rewriting strategy incurs significant latency per instance. Integrating diffusion feedback directly into the generative model could mitigate this, but real-world diffusion signals are often noisy, hindering reliable feedback integration. Our method currently does not account for temporal variations in audience features, and incorporating such dynamics remains an important direction for future research. Another future work is to explore more modalities beyond text and image. Extending the framework to audio and video could enhance its practicality.

Acknowledgments

Nan Cao is the corresponding author. This work was supported by the National Key Research and Development Program of China (2023YFB3107100).

References

Alipour, S.; Galeazzi, A.; Sangiorgio, E.; Avalle, M.; Bojic, L.; Cinelli, M.; and Quattrociocchi, W. 2024. Cross-platform social dynamics: an analysis of ChatGPT and COVID-19 vaccine conversations. *Scientific Reports*, 14(1): 2789.

Bakshy, E.; Rosenn, I.; Marlow, C.; and Adamic, L. 2012. The role of social networks in information diffusion. In *Proceedings of the international conference on World Wide Web*, 519–528.

Benevenuto, F.; Rodrigues, T.; Cha, M.; and Almeida, V. 2009. Characterizing user behavior in online social networks. In *Proceedings of the ACM SIGCOMM Conference on Internet Measurement*, 49–62.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chen, J.; Yu, J.; Ge, C.; Yao, L.; Xie, E.; Wu, Y.; Wang, Z.; Kwok, J.; Luo, P.; Lu, H.; and Li, Z. 2023. PixArt-α: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis. arXiv:2310.00426.

Chen, W.; Wang, Y.; and Yang, S. 2009. Efficient influence maximization in social networks. In *Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining*, 199–208.

Cheng, J.; Adamic, L.; Dow, P. A.; Kleinberg, J. M.; and Leskovec, J. 2014. Can cascades be predicted? In *Proceedings of the international conference on World wide web*, 925–936.

Coppolillo, E.; Cinus, F.; Minici, M.; Bonchi, F.; and Manco, G. 2025. Engagement-driven content generation with large language models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, 369–379.

De Francisci Morales, G.; Monti, C.; and Starnini, M. 2021. No echo in the chambers of political interactions on Reddit. *Scientific reports*, 11(1): 2818.

Dumoulin, V.; Belghazi, I.; Poole, B.; Mastropietro, O.; Lamb, A.; Arjovsky, M.; and Courville, A. 2017. Adversarially Learned Inference. arXiv:1606.00704.

Etta, G.; Sangiorgio, E.; Di Marco, N.; Avalle, M.; Scala, A.; Cinelli, M.; and Quattrociocchi, W. 2023. Characterizing engagement dynamics across topics on Facebook. *Plos one*, 18(6): e0286150.

Flamino, J.; Galeazzi, A.; Feldman, S.; Macy, M. W.; Cross, B.; Zhou, Z.; Serafino, M.; Bovet, A.; Makse, H. A.; and Szymanski, B. K. 2023. Political polarization of news media and influencers on Twitter in the 2016 and 2020 US presidential elections. *Nature Human Behaviour*, 7(6): 904–916.

Goldenberg, J.; Libai, B.; and Muller, E. 2001. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing letters*, 12: 211–223.

Gomez-Rodriguez, M.; Leskovec, J.; and Krause, A. 2012. Inferring networks of diffusion and influence. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 5(4): 1–37.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Granovetter, M. 1978. Threshold models of collective behavior. *American journal of sociology*, 83(6): 1420–1443.

Guille, A.; Hacid, H.; Favre, C.; and Zighed, D. A. 2013. Information diffusion in online social networks: A survey. *ACM Sigmod Record*, 42(2): 17–28.

Hodas, N. O.; and Lerman, K. 2014. The simple rules of social contagion. *Scientific reports*, 4(1): 4343.

Iamnitchi, A.; Hall, L. O.; Horawalavithana, S.; Mubang, F.; Ng, K. W.; and Skvoretz, J. 2023. Modeling information diffusion in social media: data-driven observations. *Frontiers in Big Data*, 6: 1135191.

Islam, M. R.; Muthiah, S.; Adhikari, B.; Prakash, B. A.; and Ramakrishnan, N. 2018. Deepdiffuse: Predicting the'who'and'when'in cascades. In *2018 IEEE international conference on data mining (ICDM)*, 1055–1060. IEEE.

Jiang, J.; Ren, X.; and Ferrara, E. 2023. Retweet-bert: political leaning detection using language features and information diffusion on social networks. In *Proceedings of the international AAAI conference on web and social media*, volume 17, 459–469.

Kempe, D.; Kleinberg, J.; and Tardos, É. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining*, 137–146.

Kingma, D. P.; Welling, M.; et al. 2013. Auto-encoding variational bayes.

Li, D.; Li, J.; and Hoi, S. 2023. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36: 30146–30166.

Li, Y.; and Xie, Y. 2020. Is a picture worth a thousand words? An empirical study of image content and social media engagement. *Journal of marketing research*, 57(1): 1–19.

Meng, F.; Xie, J.; Sun, J.; Xu, C.; Zeng, Y.; Wang, X.; Jia, T.; Huang, S.; Deng, Y.; and Hu, Y. 2025. Spreading dynamics of information on online social networks. *Proceedings of the National Academy of Sciences*, 122(4): e2410227122.

Notarmuzi, D.; Castellano, C.; Flammini, A.; Mazzilli, D.; and Radicchi, F. 2022. Universality, criticality and complexity of information propagation in social media. *Nature communications*, 13(1): 1308.

Pastor-Satorras, R.; and Vespignani, A. 2001. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14): 3200.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. arXiv:2307.01952.

Pressgrove, G.; McKeever, B. W.; and Jang, S. M. 2018. What is Contagious? Exploring why content goes viral on Twitter: A case study of the ALS Ice Bucket Challenge. *International Journal of Nonprofit and Voluntary Sector Marketing*, 23(1): e1586.

Prolific. 2025. Prolific: Participant recruitment for online research. https://www.prolific.com. Accessed: 2025-06-06.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Radford, A.; Metz, L.; and Chintala, S. 2016. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv:1511.06434.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

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

Sankar, A.; Zhang, X.; Krishnan, A.; and Han, J. 2020. Inf-VAE: A variational autoencoder framework to integrate homophily and influence in diffusion prediction. In *Proceedings of the international conference on web search and data mining*, 510–518.

Shahbaznezhad, H.; Dolan, R.; and Rashidirad, M. 2021. The role of social media content format and platform in users' engagement behavior. *Journal of Interactive Marketing*, 53(1): 47–65.

Spasojevic, N.; Li, Z.; Rao, A.; and Bhattacharyya, P. 2015. When-to-post on social networks. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2127–2136.

Sutton, R. S.; McAllester, D.; Singh, S.; and Mansour, Y. 1999. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12.

Tian, X.; Li, W.; Xu, B.; Yuan, Y.; Wang, Y.; and Shen, H. 2025. MIGE: Mutually Enhanced Multimodal Instruction-Based Image Generation and Editing. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, 10622–10631. ACM.

Tong, H.; Prakash, B. A.; Eliassi-Rad, T.; Faloutsos, M.; and Faloutsos, C. 2012. Gelling, and melting, large graphs by edge manipulation. In *Proceedings of the ACM international conference on Information and knowledge management*, 245–254.

Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *science*, 359(6380): 1146–1151.

Weng, L.; Menczer, F.; and Ahn, Y.-Y. 2013. Virality prediction and community structure in social networks. *Scientific reports*, 3(1): 2522.

Xu, J.; Lu, T.-C.; Compton, R.; and Allen, D. 2014. Quantifying cross-platform engagement through large-scale user alignment. In *Proceedings of the ACM conference on Web science*, 281–282.

Zhang, B.; Zhang, P.; Dong, X.; Zang, Y.; and Wang, J. 2024. Long-clip: Unlocking the long-text capability of clip. In *European conference on computer vision*, 310–325. Springer.

Appendix

A Agent Details

We provide the details of agent implementation and prompt templates we used in the information editor.

A.1 Agent for Text Generation

For text generation, we use a large language model to rewrite tweets, guided by a dynamic prompt that incorporates an action vector to adjust the six STEPPS dimensions. Given a sampled action vector $[w_1, w_2, \ldots, w_6]$, where each w_i specifies the degree of modification for the corresponding STEPPS dimension, the vector is embedded into the following prompt template:

Prompt for Text in Our Method

You are an expert in social media content crafting, adept at subtly fine-tuning text to amplify its shareability while keeping it natural and authentic.
Your task is to optimize the following social media post, making it more likely to be shared, while fully preserving the original intent, tone, and personal voice of the message.
This is a real social media post written by a human and meant to be shared by other humans. Any adjustments you make should feel seamless, as if the original author could have written them naturally.
Weight Instructions
You will optimize the post according to Jonah Berger's STEPPS framework, adjusting six specific dimensions. Each dimension has an associated weight ranging from -100% to +100%, indicating how much its influence should be reduced or enhanced:

- Positive values indicate the need to enhance the expression in the corresponding dimension (+100% means maximum enhancement).
- Negative values indicate the need to reduce the expression in the corresponding dimension (-100% means maximum reduction).
- 0% indicates no adjustment is needed for that dimension.

Please Precisely Optimize the Tweet Based on the Following STEPPS Dimensions and Their Corresponding Weights

1. Social Currency: Does the content enhance the sharer's image (e.g., look good, intelligent, funny, or like a trendsetter)? (Emphasis: {`w1`}%)
2. Triggers: Is the content strongly connected to real-life scenarios, promoting frequent recall? (Emphasis: {`w2`}%)
3. Emotion: Does the content evoke high-arousal emotions (e.g., excitement, surprise, delight)? (Emphasis: {`w3`}%)
4. Public: Is the content format easily shareable and visible? (Emphasis: {`w4`}%)
5. Practical Value: Does the content offer concrete guidelines or useful information? (Emphasis: {`w5`}%)
6. Stories: Does the content include a narrative framework or memorable elements? (Emphasis: {`w6`}%)

Optimization Requirements

1. The original meaning, tone, and personal expression style of the post must remain intact.
2. Ensure all edits blend smoothly into the text, without sounding forced or artificial.
3. Maintain the original language of the post (e.g., if it's in English, the output must be in English).
4. The optimized text must have visible adjustments, and cannot be identical to the original text.
5. The optimized version should feature diverse sentence structures and vivid, flexible expressions, avoiding monotony or rigid patterns.
6. Ensure the language flows smoothly with clear logic and highlighted key points, making the message more impactful.
7. Where appropriate, you may add emojis to enhance emotional resonance and visual appeal, as long as they align with the original tone.
8. Remember, this is a social media post that should feel authentic, engaging, and human.

Output Format Requirements

1. The output MUST be in JSON format with the following fields:
   - `original_text`: The exact original input text.
   - `optimized_text`: The human-like, optimized version of the text.
2. Return ONLY the JSON object, without any additional explanations or formatting.

A.2 Agent for Image Generation

Similarity, for image generation, we sample an action vector $[v_1, v_2, \ldots, v_8]$, where each component corresponds to one of the eight perceptual attributes associated with user engagement. This vector is embedded in the following dynamic prompt that guides the rewriting of the text-to-image input. The resulting prompt is then fed into an image generation model to produce the final image.

Prompt for Image in Our Method

You are a text-to-image prompt engineer, skilled in drafting and precisely optimizing prompts to maximize the visual appeal and virality of generated images on social media platforms.
Your task is to refine the provided text-to-image prompt by rewriting it in a way that fully retains the semantic content of the original tweet, while enriching background details, ensuring the image composition is vibrant, meaningful, and aesthetically pleasing.

Weight Instructions
The following visual dimensions have been assigned specific weights. The weight, expressed as a percentage, indicates the intensity of optimization for the corresponding dimension, with a range from -100% to 100%:
- Positive values indicate the need to enhance the expression in the corresponding dimension (+100% means maximum enhancement).
- Negative values indicate the need to reduce the expression in the corresponding dimension (-100% means maximum reduction).
- 0% indicates no adjustment is needed for that dimension.

Please optimize the text-to-image prompt based on the following dimensions and their corresponding weights
1. Colorfulness: Whether the image is rich in color and has visual impact (Emphasis: {`v1`}%)
2. Human Scene: Whether the image includes scenes with people to enhance emotional resonance (Emphasis: {`v2`}%)
3. Emotion: Whether the image can evoke strong emotional responses (such as joy, shock, or being moved) (Emphasis: {`v3`}%)
4. Professional: Whether the image has the quality of professional photography (Emphasis: {`v4`}%)
5. Brightness: Whether the image is bright and attention-grabbing (Emphasis: {`v5`}%)
6. Clarity: Whether the image is clear and its details are prominent (Emphasis: {`v6`}%)
7. Visual Balance: Whether the visual elements in the image are evenly distributed (Emphasis: {`v7`}%)
8. Focus of the Picture: Whether the image focuses on the subject, avoiding visual dispersion (Emphasis: {`v8`}%)

Optimization Requirements
1. The optimized prompt must faithfully preserve the original tweet's semantics and intended message.
2. Enrich the background context and visual storytelling, ensuring the scene feels alive and dynamic.
3. Emphasize the visual vitality and aesthetic beauty of the image through natural and seamless prompt enhancements.
4. The optimized prompt MUST NOT be identical to the original prompt; rewrite with nuanced adjustments.

Output Format Requirements
1. The output MUST be in JSON format with the following fields:
   - `original_tweet_text`: The exact original input tweet text.
   - `original_prompt`: The original input text-to-image prompt before optimization.
   - `optimized_prompt`: The refined prompt after optimization, keeping within 75 words.
2. Return ONLY the JSON object, without any extra explanations or formatting.

B Evaluation Details

We add the implementation of evaluation, including more details of experimental setups and metric calculations.

B.1 Experimental Setup

We provide a more detailed description of the experimental setups for both the influence indicator and the information editor components of our framework.

Influence Indicator For the influence indicator evaluation, all experiments were conducted on a server with dual Intel Xeon Gold 6148 CPUs (40 cores) and 1 × NVIDIA Tesla V100 16GB GPU, running Ubuntu 16.04 LTS, Python 3.9.12, PyTorch 1.11.0, and CUDA 11.6. The pairwise diffusion estimator was trained for 100 epochs and a batch size of 256. The architecture consists of a shared feature extractor with two fully connected layers and a classifier with four layers, each followed by ReLU activation and dropout (rate = 0.3).

Information Editor For the information editor evaluation, all experiments were conducted on a server with dual Intel Xeon Gold 6348 CPUs (40 cores) and 2 × NVIDIA A800 80GB GPUs, running Ubuntu 20.04 LTS, Python 3.8.13, and CUDA 12.2. Text generation utilized Azure's GPT-4 API, while image generation was performed using a locally deployed PixArt-α model, configured with 30 denoising inference steps per sample. The policy networks for text and image editing were trained separately for 350 episodes, each comprising three iterative rewriting steps, using the Adam optimizer with a learning rate of 1×10^{-4}. During inference, the policy networks apply the same three-step rewriting process and select the message yielding the highest reward.

B.2 Metric Calculation

We assess our information editor using two metrics:
1. *Diffusion Gain*, which measures the relative increase of influence scores after rewriting, and defined as

$$\text{Diffusion Gain} = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathcal{L}_U(c_i') - \mathcal{L}_U(c_i)}{\mathcal{L}_U(c_i)} \times 100\% \tag{1}$$

   where N is the number of content instances, c_i and c_i' are the original and edited messages, respectively;
2. *Consistency*, which quantifies semantic preservation between original and edited content via the cosine similarity of their features.

$$\text{Consistency} = \frac{1}{N}\sum_{i=1}^{N}\cos\left(f(c_i), f(c_i')\right) \tag{2}$$

B.3 Baseline Implementation

We provide the details of baseline implementation.

Large Language Model (LLM) The Large Language Model (LLM) baseline performs a one-step rewrite of input texts —-such as tweets or text-to-image prompts—without employing task-specific editing strategies or reference guidance. It uses a general-purpose prompt aimed at generating high-quality rewrites. Unlike our method, this baseline does not incorporate the designed action space, which defines editing dimensions and corresponding weight instructions for fine-grained, controllable edits. Apart from this, all other settings, including the prompt structure, remain identical to those in our method. The baseline prompt is constructed as follows:

Prompt for Text in LLM

You are an expert in social media content crafting, adept at subtly fine-tuning text to amplify its shareability while keeping it natural and authentic.
Your task is to optimize the following social media post, making it more likely to be shared, while fully preserving the original intent, tone, and personal voice of the message.
This is a real social media post written by a human and meant to be shared by other humans. Any adjustments you make should feel seamless, as if the original author could have written them naturally.
Optimization Requirements
1. The original meaning, tone, and personal expression style of the post must remain intact.
2. Ensure all edits blend smoothly into the text, without sounding forced or artificial.
3. Maintain the original language of the post (e.g., if it's in English, the output must be in English).
4. The optimized text must have visible adjustments, and cannot be identical to the original text.
5. The optimized version should feature diverse sentence structures and vivid, flexible expressions, avoiding monotony or rigid patterns.
6. Ensure the language flows smoothly with clear logic and highlighted key points, making the message more impactful.
7. Where appropriate, you may add emojis to enhance emotional resonance and visual appeal, as long as they align with the original tone.
8. Remember, this is a social media post that should feel authentic, engaging, and human.

Output Format Requirements
1. The output MUST be in JSON format with the following fields:
   - `original_text`: The exact original input text.
   - `optimized_text`: The human-like, optimized version of the text.
2. Return ONLY the JSON object, without any additional explanations or formatting.

Prompt for Image in LLM

You are a text-to-image prompt engineer, skilled in drafting and precisely optimizing prompts to maximize the visual appeal and virality of generated images on social media platforms.
Your task is to refine the provided text-to-image prompt by rewriting it in a way that fully retains the semantic content of the original tweet, while enriching background details, ensuring the image composition is vibrant, meaningful, and aesthetically pleasing.
Optimization Requirements
1. The optimized prompt must faithfully preserve the original tweet's semantics and intended message.
2. Enrich the background context and visual storytelling, ensuring the scene feels alive and dynamic.
3. Emphasize the visual vitality and aesthetic beauty of the image through natural and seamless prompt enhancements.
4. The optimized prompt MUST NOT be identical to the original prompt; rewrite with nuanced adjustments.

Output Format Requirements
1. The output MUST be in JSON format with the following fields:
   - `original_tweet_text`: The exact original input tweet text.
   - `original_prompt`: The original input text-to-image prompt before optimization.
   - `optimized_prompt`: The refined prompt after optimization, keeping within 75 words.
2. Return ONLY the JSON object, without any extra explanations or formatting.

In-Context Learning (IC-L) The In-Context Learning (IC-L) baseline performs a one-step rewrite by selecting, for each audience group, messages with the highest and lowest observed influence scores as positive and negative exemplars, respectively. These exemplars are used as few-shot prompts to guide LLM-based rewriting. Apart from this, all other settings remain identical to the standard LLM baseline. The IC-L prompts are constructed as follows:

Prompt for Text in IC-L

You are an expert in social media content crafting, adept at subtly fine-tuning text to amplify its shareability while keeping it natural and authentic.
Your task is to optimize the following social media post, making it more likely to be shared, while fully preserving the original intent, tone, and personal voice of the message.
This is a real social media post written by a human and meant to be shared by other humans. Any adjustments you make should feel seamless, as if the

original author could have written them naturally.
You will be provided with several examples of tweet messages:
- 3 positive exemplars with high diffusion effectiveness.
- 3 negative exemplars with low diffusion effectiveness.

Your goal is to analyze the difference between these examples and rewrite the input message to match the tone, structure, and virality potential of the positive exemplars while avoiding the issues present in the negative ones.
Optimization Requirements
1. The original meaning, tone, and personal expression style of the post must remain intact.
2. Ensure all edits blend smoothly into the text, without sounding forced or artificial.
3. Maintain the original language of the post (e.g., if it's in English, the output must be in English).
4. The optimized text must have visible adjustments, and cannot be identical to the original text.
5. The optimized version should feature diverse sentence structures and vivid, flexible expressions, avoiding monotony or rigid patterns.
6. Ensure the language flows smoothly with clear logic and highlighted key points, making the message more impactful.
7. Where appropriate, you may add emojis to enhance emotional resonance and visual appeal, as long as they align with the original tone.
8. Remember, this is a social media post that should feel authentic, engaging, and human.

Output Format Requirements
1. The output MUST be in JSON format with the following fields:
   - `original_text`: The exact original input text.
   - `optimized_text`: The human-like, optimized version of the text.
2. Return ONLY the JSON object, without any additional explanations or formatting.

Positive Exemplars (High Diffusion)
{positive_exemplars}
Negative Exemplars (Low Diffusion)
{negative_exemplars}

Prompt for Image in IC-L

You are a text-to-image prompt engineer, skilled in drafting and precisely optimizing prompts to maximize the visual appeal and virality of generated images on social media platforms.
Your task is to refine the provided text-to-image prompt by rewriting it in a way that fully retains the semantic content of the original tweet, while enriching background details, ensuring the image composition is vibrant, meaningful, and aesthetically pleasing.
You will be provided with several examples:
- 3 positive exemplars where the prompt led to highly viral images.
- 3 negative exemplars where the prompt led to low-performing images.

Your goal is to analyze what makes the positive examples more visually engaging and viral, and then apply similar improvements to the current prompt while maintaining the original semantics and visual style.

Optimization Requirements
1. The optimized prompt must faithfully preserve the original tweet's semantics and intended message.
2. Enrich the background context and visual storytelling, ensuring the scene feels alive and dynamic.
3. Emphasize the visual vitality and aesthetic beauty of the image through natural and seamless prompt enhancements.
4. The optimized prompt MUST NOT be identical to the original prompt; rewrite with nuanced adjustments.

Output Format Requirements
1. The output MUST be in JSON format with the following fields:
   - `original_tweet_text`: The exact original input tweet text.
   - `original_prompt`: The original input text-to-image prompt before optimization.
   - `optimized_prompt`: The refined prompt after optimization, keeping within 75 words.
2. Return ONLY the JSON object, without any extra explanations or formatting.

Positive Exemplars (High Diffusion)
{positive_exemplars}
Negative Exemplars (Low Diffusion)
{negative_exemplars}

Greedy Search (Greedy) The Greedy Search (Greedy) baseline is a non-learning variant of our framework that uniformly samples editing actions from the action space $[-1, 1]^n$ at each step. It performs three iterative rewriting steps and selects the message with the highest reward. The only difference from our method is that Greedy samples action vectors uniformly, while our method samples them from the policy network's learned distribution.

B.4 User Study Details

We provide additional details on participant recruitment and survey design.

Participant Recruitment We recruited participants who met the following criteria: native English speakers, X users, and interested in sports. All participants gave informed con-

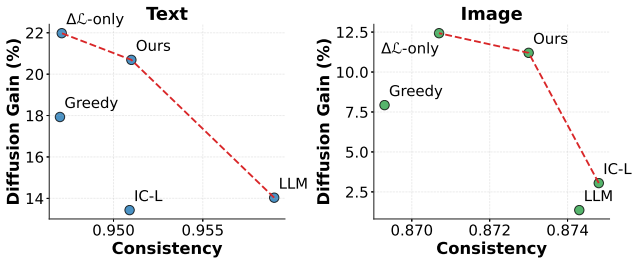

Figure 1: Pareto-style visualization of diffusion gain versus consistency across methods for both text and image tasks.

sent prior to the study, and no personally identifiable information was collected.

Survey Design The survey consists of two parts: a brief background information collection and a content preference task. For the text task, participants were presented with pairs of tweets and asked, *"When you see these two tweets on Twitter (X), which one would you be more willing to share/repost?"* For the image task, participants viewed pairs of images accompanied by a tweet prompt and were asked, *"Which of the following images would you be more willing to share/repost along with the tweet above?"* Pairs were randomized to minimize order effects. After completing tasks, they selected reasons for their preferences from a multiple-choice list.

C Pareto Front Visualization

To illustrate the trade-off between diffusion gain and semantic consistency discussed in the ablation study, we include a Pareto-style visualization in Figure 1. The figure plots all evaluated methods on the two-dimensional plane of diffusion gain versus consistency, with the dashed red line connecting the non-dominated (Pareto-optimal) points. Our method lies near the Pareto front, indicating a balanced trade-off between diffusion gain and semantic consistency.

D Ethical Concerns

Our framework is designed to study and model information diffusion, but its ability to enhance content virality raises several ethical considerations. First, the editing mechanism may unintentionally amplify manipulative or misleading content if applied without proper safeguards. To mitigate this, all experiments are conducted on publicly available, anonymized data, and the system is restricted to style-level rewriting without altering factual claims. Second, controllable editing introduces risks regarding user autonomy, as optimized content could influence sharing behavior in subtle ways. We emphasize that the method is intended for research on diffusion mechanisms rather than real-world deployment. Finally, both text and image generation comply with platform safety filters to prevent harmful, hateful, or deceptive outputs. Future work should further integrate robust content moderation and human oversight mechanisms.