AgentBook: Enabling Fine-Grained Editability in Visual Narratives via Page-Centric Agent Coordination

Ayushman Sarkar[1], Zhenyu Yu[2]
[1]*Birbhum Institute of Engineering and Technology (Affiliated to MAKAUT)*, Suri, West Bengal, 731101, India
[2]*Lincoln University College*, Kuala Lumpur 50603, Malaysia
ayushmansarkar123@gmail.com, yuzhenyuyxl@foxmail.com

***Abstract*—Long-horizon visual storytelling with text-to-image diffusion models enables coherent multi-page narratives from a single prompt, yet existing single-pass generation pipelines tightly couple pages within a shared latent trajectory, limiting structured editability. We propose AgentBook, a training-free multi-agent framework that reformulates story generation as a page-centric process governed by an explicit global narrative state encoding character identity, stylistic constraints, and narrative context. By decoupling page synthesis from a monolithic diffusion chain and coordinating story planning, generation, consistency enforcement, and user editing through state-based conditioning, AgentBook enables localized page regeneration and iterative refinement without compromising cross-page coherence. Experimental results demonstrate improvements in story coherence, visual consistency, and edit locality over representative visual storytelling and text-to-image diffusion baselines, establishing a controllable and interactive paradigm for long-form visual narrative generation.**

***Index Terms*—Visual storytelling, text-to-image diffusion, multi-agent systems, interactive generation, cross-page consistency, editable narrative synthesis.**

I. Introduction

Text-to-image (T2I) diffusion models [1], [2] have significantly advanced high-fidelity visual generation, enabling controllable synthesis from natural language prompts. Recent developments in diffusion-based generative modeling [3], [4] further improve semantic alignment, identity preservation, and stylistic consistency. Building on these advances, long-horizon storytelling frameworks such as One-Prompt-One-Story (1Prompt1Story) [5] demonstrate that coherent multi-page visual narratives can be generated from a single prompt without retraining.

Despite these advances, existing approaches typically rely on one-pass generation under a shared diffusion trajectory, tightly coupling all pages. While this ensures overall coherence, it limits structured editability, as modifying a single page often requires regenerating the entire story. In contrast, real-world story creation is iterative and page-centric, where individual pages are refined without reconstructing the full narrative.

To address this limitation, we reformulate visual storytelling as a structured, page-level generation problem governed by a shared global narrative state. Each page is generated independently while maintaining global constraints on character identity, style, and narrative progression, enabling localized edits without disrupting overall coherence.

We propose **AgentBook**, a framework for editable visual storytelling. AgentBook decomposes story generation into coordinated components for planning, page synthesis, consistency enforcement, and user editing. By decoupling page generation from a single trajectory while maintaining state-based coordination, our approach supports localized editing, iterative refinement, and human-in-the-loop control while preserving cross-page coherence.

The main **contributions** are summarized as follows:

- **Structural Reformulation of Long-Horizon Storytelling.** We introduce a modular page-based abstraction that addresses the entangled generation constraints of single-pass diffusion pipelines, aligning automated storytelling with iterative, page-centric workflows.
- **State-Governed Multi-Agent Coordination.** We present a training-free, multi-agent framework that generates pages independently while enforcing global narrative consistency, explicitly supporting controllable user interventions.
- **Editable Storybook Evaluation Protocol.** We develop quantitative and human-centered metrics to evaluate page-level editability, narrative coherence retention, and user control in interactive storytelling scenarios.

II. Related Work

A. Visual Storytelling and Long-Horizon Generation

Recent advances in text-to-image diffusion models [1]–[4] have enabled high-quality image synthesis from textual descriptions, making them suitable for visual storytelling. However, maintaining character identity, stylistic coherence, and narrative consistency across multiple images remains challenging in long-horizon generation. Prior works such as StoryDALL-E [6], StoryGen [7], and 1Prompt1Story [5] explore diffusion-based story generation with improved cross-frame consistency. Meanwhile, general-purpose models (e.g., PixArt-α [8], Playground-v2 [9]) can generate high-quality sequential images but typically rely on tightly coupled pipelines, limiting modular reasoning and flexible page-level editing.

B. Human-in-the-Loop and Interactive Generation

Interactive generative systems emphasize iterative refinement and user-guided content creation [10], [11]. These approaches enable users to progressively guide generation through prompt refinement, attention guidance, or mixed-initiative interaction. While such techniques improve controllability in single-image generation, most current storytelling frameworks do not directly integrate user edits while maintaining cross-page narrative consistency in multi-image stories.

C. Agent-Based Generative Systems

Recent research has explored agent-based frameworks that decompose complex generative tasks into coordinated modules [12], [13]. Multi-agent systems introduce planning, coordination, and reasoning mechanisms that enable structured generation pipelines. However, existing agent-based frameworks have not been specifically designed for long-horizon visual storytelling, where character identity, narrative state, and page-level edits must remain consistent across multiple generated images.

Our approach, **AgentBook**, introduces a state-governed multi-agent architecture for visual story generation. Unlike prior monolithic pipelines, AgentBook models story generation as a structured process involving planning, rendering, consistency monitoring, and user-driven editing operating over an explicit global narrative state. This design enables modular story construction, localized edits, and improved cross-page consistency.

III. Agentic Storybook Generation Framework

We formulate visual storytelling as a *state-conditioned page-level generation* problem and introduce AgentBook, a structured framework that enables modular synthesis with explicit editability. Instead of binding all pages within a single diffusion trajectory, we factorize story generation into page-wise processes conditioned on a shared global narrative state.

A. Problem Formulation

We represent a storybook as an ordered sequence of semantic pages:

$$\mathcal{B} = \{(s_i, I_i)\}_{i=1}^{N}, \tag{1}$$

where s_i denotes the semantic description and I_i the generated image of page i. To enforce cross-page consistency, we introduce a global narrative state:

$$S = \{c_1, c_2, ..., c_K, \theta_{style}, \theta_{story}\}, \tag{2}$$

where c_k represents persistent character attributes, θ_{style} encodes stylistic constraints, and θ_{story} captures narrative context. We model S as a structured graph, where nodes correspond to characters and scenes, and edges encode relational constraints such as co-occurrence and narrative dependency.

The generation objective is to synthesize each page conditioned on both its semantics and the global state:

$$I_i \sim G(s_i, S), \tag{3}$$

where $G(\cdot)$ denotes the diffusion-based image generator. In practice, conditioning is implemented via guided denoising:

$$x_{t-1} = \mu_\theta(x_t, s_i, S) + \sigma_t \epsilon, \tag{4}$$

where μ_θ is the conditional denoising network and $\epsilon \sim \mathcal{N}(0, I)$.

To enforce character consistency across pages, we define a global coherence objective:

$$\mathcal{L}_{cons} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \sum_{i,j} d\big(f_c(I_i), f_c(I_j)\big), \tag{5}$$

where $f_c(\cdot)$ extracts character-specific features and $d(\cdot, \cdot)$ measures feature similarity. During interactive editing, a user instruction E modifies page semantics:

$$s_t' = \Phi(s_t, E), \tag{6}$$

where $\Phi(\cdot)$ maps natural language instructions to structured constraints. The edited page is then regenerated as:

$$I_t' \sim G(s_t', S). \tag{7}$$

B. Storybook Representation and Editability

Unlike video generation, which models temporally continuous frames, a storybook consists of discrete semantic pages, each representing a complete narrative unit. This structure allows page-level independence while requiring explicit mechanisms to enforce cross-page consistency in character identity, style, and narrative progression. Such a formulation naturally supports localized editing, where individual pages can be regenerated without affecting the entire story, while maintaining global coherence through shared constraints.

C. AgentBook Architecture

We model story generation as a factorized process over pages conditioned on a shared global state:

$$P(\mathcal{I}_{1:N} \mid S) = \prod_{i=1}^{N} P(I_i \mid S, C_i), \tag{8}$$

where C_i denotes the page-level semantic description.

To operationalize this formulation, AgentBook decomposes the generation process into four coordinated components interacting through the global state S.

Story Planning. Generates structured page-level descriptions $\{C_i\}_{i=1}^{N}$ from a high-level prompt. **Page Generation.** Synthesizes each page independently using a diffusion model conditioned on (C_i, S). **User Editing.** Transforms natural language instructions into structured constraints and determines whether updates are local or global. **Consistency Enforcement.** Maintains cross-page identity, style, and narrative coherence via updates to the global state.

To enforce localized editing, we introduce an editing constraint:

$$\mathcal{L}_{edit} = \sum_{i \neq t} d\big(F(I_i), F(I_i')\big), \tag{9}$$

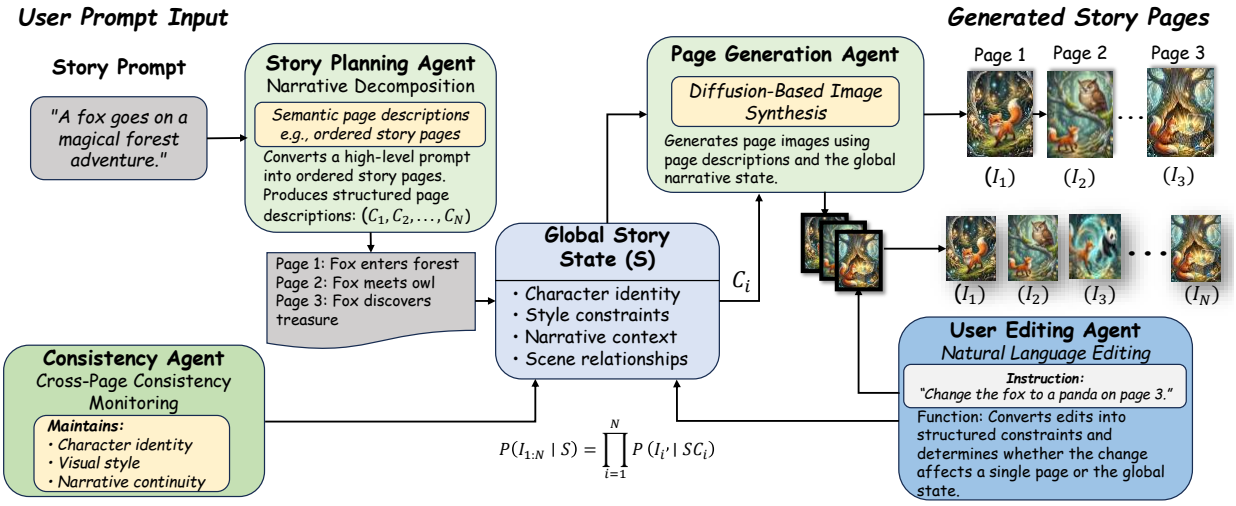

Fig. 1. Overview of the proposed AgentBook framework. The system decomposes story generation into planning, page synthesis, consistency enforcement, and user-guided editing, coordinated through a global narrative state.

where $F(\cdot)$ denotes a feature extractor (e.g., CLIP) and $d(\cdot, \cdot)$ is a similarity distance. The global state is updated using features extracted from generated pages:

$$S_{i+1} = \Psi(S_i, f(I_i)), \tag{10}$$

where $f(I_i)$ extracts semantic attributes and $\Psi(\cdot)$ integrates them into the state.

This update can be interpreted as a state-space process:

$$S_{i+1} \sim P(S_{i+1} \mid S_i, I_i), \quad I_i \sim P(I_i \mid S_i, C_i). \tag{11}$$

IV. Experiments

A. Experimental Setup and Baselines

Dataset. To ensure fair comparison with prior story-consistent text-to-image generation methods, we adopt the public *ConsiStory+* benchmark [5], which has been used in recent studies on multi-image narrative consistency. The dataset provides structured prompt sequences describing characters, environments, and sequential events within a story. Following the standard setup, each story contains $N = 6$ pages with corresponding semantic descriptions. To increase prompt diversity and narrative clarity, we slightly refine the original prompt sequences using GPT-4o while preserving the core semantic structure of the dataset.

Implementation Details. AgentBook employs a Story Planning Agent that decomposes high-level prompts into structured page semantics $\{s_i\}_{i=1}^{N}$ while maintaining a global story state S encoding character attributes, scene descriptors, and narrative context. The Page Generation Agent synthesizes page images conditioned on (s_i, S) using Stable Diffusion v1.5 as the underlying text-to-image diffusion backbone, while a Consistency Agent monitors cross-page visual alignment. For interactive tasks, a User-Editing Agent enables targeted page modifications without regenerating the entire story sequence.

Baselines. We compare AgentBook against representative visual storytelling and diffusion-based image generation models including StoryDALL-E, StoryGen, PixArt-α, Playground-v2, and 1Prompt1Story. All baseline models are executed using the default configurations provided by their respective publications or official implementations. For fair comparison, identical prompts and inference settings are used across all models when generating multi-page story sequences.

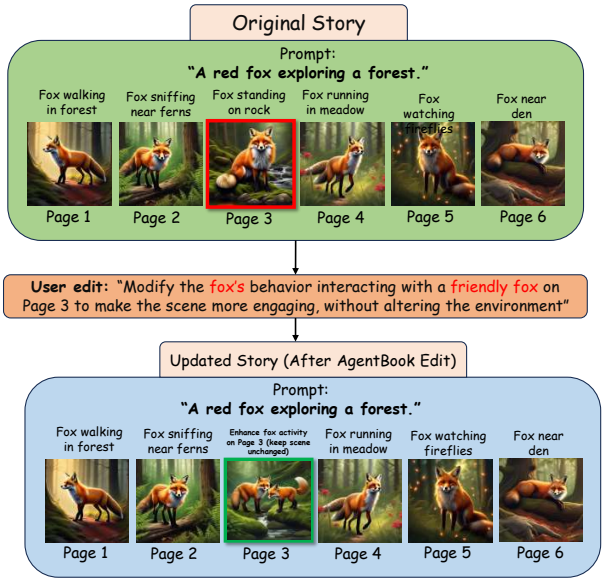

Fig. 2. Localized editing demonstration. Given a coherent multi-page story, AgentBook modifies only the target page based on user instruction (e.g., Modify the fox's behavior interacting with a friendly fox on Page 3 to make the scene more engaging, without altering the environment), while preserving all other pages unchanged, maintaining global narrative and visual consistency.

Evaluation Metrics. We evaluate AgentBook using both automatic metrics and human studies, focusing on semantic alignment, cross-page consistency, and edit locality. Specifically, CLIP-based text alignment (CLIP-T) measures the consistency between generated images and their corresponding

descriptions [14], [15], while CLIP-based image similarity (CLIP-I) evaluates visual consistency across pages containing recurring characters or scenes [2], [7]. To assess editing precision, we adopt the Editing Locality Ratio (ELR), which quantifies the extent to which user edits remain confined to the target page without affecting others, computed as

$$ELR = \frac{1}{N-1} \sum_{i \neq t} (1 - \text{sim}\left(F(I_i), F(I_i')\right)) \qquad (12)$$

where N denotes the total number of pages, t represents the edited target page, I_i and I_i' denote the original and post-edit versions of page i, respectively, $F(\cdot)$ represents the feature extraction function, and $\text{sim}(\cdot, \cdot)$ denotes cosine similarity between extracted features; lower ELR values indicate better editing locality since smaller changes outside the target page correspond to stronger edit containment.

For editing evaluation, we apply controlled modifications (e.g., character changes, object insertion, and scene edits) on randomly selected pages to test both locality and global consistency. In addition, we conduct a human study with 100 participants over 80 sampled stories, where outputs are evaluated via blind pairwise comparison based on narrative coherence, character consistency, and editing quality, following established human evaluation protocols.

B. Quantitative Results

Table I presents a quantitative comparison with representative baselines. AgentBook achieves the best performance on both semantic alignment and visual consistency, with CLIP-T (0.912) and CLIP-I (0.884), outperforming the strongest baseline (1Prompt1Story) by +0.038 and +0.067, respectively. In terms of efficiency, AgentBook builds on the single-pass 1Prompt1Story [5] generation pipeline and introduces only modest additional overhead from page-wise agent coordination; its per-story inference time therefore remains close to, and slightly higher than, that of 1Prompt1Story. We regard this small overhead as a worthwhile trade-off, since the page-centric design enables localized editing without regenerating the entire story.

In contrast, existing approaches still exhibit trade-offs between alignment, consistency, and editability. While StoryGen achieves relatively strong CLIP-I, it underperforms in semantic alignment. General-purpose models (e.g., PixArt-α, Playground-v2) produce high-quality individual images but struggle to maintain consistent character identity across pages due to the lack of explicit global state modeling. Fig. 3 further illustrates that AgentBook produces more consistent character identity and coherent narrative progression across multi-page story sequences.

C. Ablation Study

We conduct an ablation study to evaluate the contribution of each core component in AgentBook, including the Story Planning Agent, Consistency Agent, and User Editing Agent. Each component is removed independently under the same experimental setting using CLIP-T, CLIP-I, and ELR. As shown in Table II, removing the **Story Planning Agent** reduces CLIP-T from 0.912 to 0.861, indicating degraded semantic alignment. Removing the **Consistency Agent** leads to the largest drop in CLIP-I (0.884 to 0.791), confirming its role in maintaining cross-page visual consistency. Removing the **User Editing Agent** significantly increases ELR (0.12 to 0.74), indicating poor edit locality with unintended changes propagating to non-target pages. Overall, the full model achieves the best performance across all metrics, demonstrating that structured planning, global consistency modeling, and edit-aware generation jointly contribute to coherent and controllable story generation.

Fig. 3. Comparison with representative story generation models across multi-page narratives. AgentBook produces more consistent character identity and coherent story progression compared to prior approaches.

TABLE I
QUANTITATIVE COMPARISON OF AGENTBOOK WITH REPRESENTATIVE VISUAL STORYTELLING AND TEXT-TO-IMAGE BASELINES ON 100 PROMPTS FROM THE CONSISTORY+ BENCHMARK. HIGHER IS BETTER (↑) FOR BOTH CLIP-T AND CLIP-I. **BOLD** AND UNDERLINED INDICATE THE BEST AND SECOND-BEST RESULTS, RESPECTIVELY.

Method	CLIP-T ↑	CLIP-I ↑
StoryDALL-E [6]	0.742 ± .04	0.685 ± .05
StoryGen [7]	0.795 ± .03	0.762 ± .03
PixArt-α [8]	0.828 ± .02	0.714 ± .04
Playground-v2 [9]	0.841 ± .02	0.738 ± .03
1Prompt1Story [5]	0.874 ± .01	0.817 ± .02
AgentBook (Ours)	**0.912 ± .01**	**0.884 ± .01**

TABLE II
ABLATION STUDY ON AGENTBOOK COMPONENTS. ELR MEASURES EDITING LOCALITY (LOWER IS BETTER).

Variant	CLIP-T ↑	CLIP-I ↑	ELR ↓
w/o Story Planning	0.861	0.802	0.14
w/o Consistency Agent	0.884	0.791	0.18
w/o User Editing Agent	0.892	0.835	0.74
Full Model	**0.912**	**0.884**	**0.12**

D. User Study

We conduct a user study with 100 participants to evaluate the perceptual quality of generated storybooks. A total of

TABLE III

CAPABILITY COMPARISON OF REPRESENTATIVE VISUAL STORYTELLING AND TEXT-TO-IMAGE GENERATION APPROACHES. ✓ INDICATES THE CAPABILITY IS FULLY SUPPORTED, × INDICATES IT IS NOT SUPPORTED, AND △ DENOTES PARTIAL OR INDIRECT SUPPORT.

Method	Page-Level Editing	Character Consistency	Style Consistency	Global Story Planning	Localized Regeneration	Train-Free
StoryDALL-E [6]	×	✓	✓	×	×	×
StoryGen [7]	×	✓	✓	×	×	×
PixArt-α [8]	×	×	✓	×	×	△
Playground-v2 [9]	×	×	✓	×	×	△
1Prompt1Story [5]	×	✓	✓	△	×	✓
AgentBook (Ours)	✓	✓	✓	✓	✓	✓

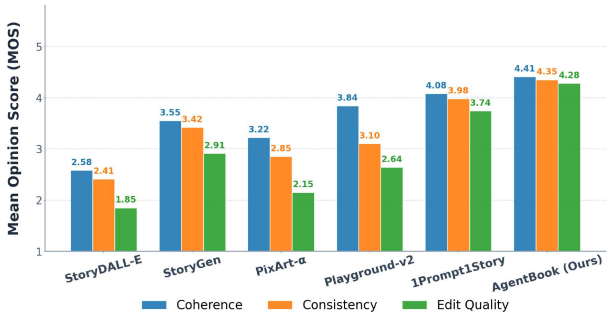

Fig. 4. User study results in terms of identity story coherence, character consistency, and editing quality.

80 stories were randomly selected, and outputs from five methods (StoryDALL-E, StoryGen, PixArt-α, Playground-v2, 1Prompt1Story and AgentBook) were presented in randomized order to reduce bias. Participants rated each method on a 1–5 Likert scale based on (i) story coherence (Q1), (ii) character consistency (Q2), and (iii) editing quality (Q3), along with an overall preference score. Results show that AgentBook outperforms all baselines across all criteria, with the largest gains observed in editing quality. As shown in Fig. 4, AgentBook is preferred in 71%, 74%, and 82% of comparisons for story coherence, character consistency, and editing quality, respectively, indicating strong localized editing capability while maintaining global narrative coherence.

V. CONCLUSION

We presented **AgentBook**, an agentic storybook generation framework that enables modular page-level synthesis with explicit editability. The framework allows creators to iteratively refine narrative, characters, and scenes while maintaining cross-page coherence. By decomposing story creation into structured agents, namely the Story Planning Agent, Page Generation Agent, User Editing Agent, and Consistency Agent, AgentBook preserves long-horizon visual and narrative consistency and supports localized edits without reconstructing the entire story. Extensive experiments and user evaluations demonstrate superior performance in coherence, cross-page consistency, and edit controllability compared to existing baselines. These results highlight the effectiveness of modular, state-aware agent coordination for controllable and interactive visual storytelling.

REFERENCES

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems (NeurIPS)*, vol. 33, pp. 6840–6851, 2020.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2022, pp. 10 684–10 695.

[3] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations (ICLR)*, 2021.

[4] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proceedings of the 38th International Conference on Machine Learning*, 2021, pp. 8162–8171. [Online]. Available: https://proceedings.mlr.press/v139/nichol21a.html

[5] T. Liu, K. Wang, S. Li, J. Van de Weijer, F. Khan, S. Yang, Y. Wang, J. Yang, and M.-M. Cheng, "One-prompt-one-story: Free-lunch consistent text-to-image generation using a single prompt," in *International Conference on Learning Representations*, vol. 2025, 2025, pp. 24 470–24 497.

[6] A. Maharana, D. Hannan, and M. Bansal, "Storydall-e: Adapting pretrained text-to-image transformers for story continuation," in *European conference on computer vision*. Springer, 2022, pp. 70–87.

[7] C. Liu, H. Wu, Y. Zhong, X. Zhang, Y. Wang, and W. Xie, "Intelligent grimm-open-ended visual storytelling via latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 6190–6200.

[8] J. Chen, J. Yu, C. Ge, L. Yao, E. Xie, Z. Wang, J. Kwok, P. Luo, H. Lu, and Z. Li, "Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis," in *International conference on learning representations (ICLR)*, vol. 2024, 2024, pp. 57 611–57 640.

[9] D. Li, A. Kamko, E. Akhgari, A. Sabet, L. Xu, and S. Doshi, "Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation," *arXiv preprint arXiv:2402.17245*, 2024.

[10] F. Faltings, M. Galley, K. Brantley, B. Peng, W. Cai, Y. Zhang, J. Gao, and W. B. Dolan, "Interactive text generation," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 4450–4468.

[11] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[12] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," in *Proceedings of the 36th annual acm symposium on user interface software and technology*, 2023, pp. 1–22.

[13] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. R. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," in *The eleventh international conference on learning representations*, 2022.

[14] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning (ICML)*. PmLR, 2021, pp. 8748–8763.

[15] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," in *Proceedings of the 2021 conference on empirical methods in natural language processing*, 2021, pp. 7514–7528.