AcT2I: Evaluating and Improving Action Depiction in Text-to-Image Models

Anonymous ACL submission

Abstract

Text-to-Image (T2I) models have recently achieved remarkable success in generating images from textual descriptions. However, challenges still persist in accurately rendering complex scenes where actions and interactions form the primary semantic focus. Our key observation in this work is that T2I models frequently struggle to capture nuanced and often implicit attributes inherent in action depiction, leading to generating images that lack key contextual details. To enable systematic evaluation, we introduce AcT2I, a benchmark designed to evaluate the performance of T2I models in generating images from action-centric prompts. We experimentally validate that leading T2I models do not fare well on AcT2I. We further hypothesize that this shortcoming arises from the incomplete representation of the inherent attributes and contextual dependencies in the training corpora of existing T2I models. We build upon this by developing a training-free, knowledge distillation technique utilizing Large Language Models to address this limitation. Specifically, we enhance prompts by incorporating dense information across three dimensions, observing that injecting prompts with temporal details significantly improves image generation accuracy, with our best model achieving an increase of 72%. Our findings highlight the limitations of current T2I methods in generating images that require complex reasoning and demonstrate that integrating linguistic knowledge in a systematic way can notably advance the generation of nuanced and contextually accurate images.

1 Introduction

Text-to-Image (T2I) models have advanced rapidly, evolving from simple image generation systems to producing intricate, photorealistic scenes (Karras et al., 2019; Rombach et al., 2022; Esser et al., 2024). There has been a consistent growth in the performance of T2I models in their ability to perform compositional tasks such as object placement and attribute binding, as evidenced by their performance on benchmarks such as T2I-CompBench (Huang et al., 2023) and GenEval (Ghosh et al., 2024).

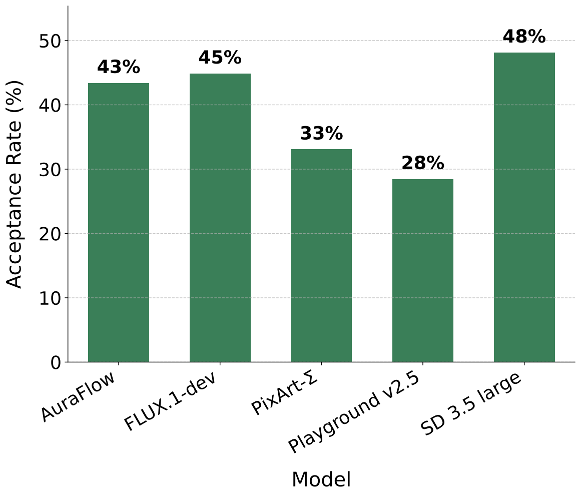


Figure 1: Action Depiction Performance of SOTA Text-to-Image Models. Each generated image was evaluated by three reviewers, resulting in a binary acceptance (yes/no). The acceptance rate represents the average agreement among reviewers on each model's acceptability of action depiction.

However, these benchmarks aim to capture *1-hop* capabilities of T2I models, i.e. evaluating abilities that does not require nuanced reasoning while generating an image. For example, to generate a "*blue apple*", the model has to follow 2 steps: **1**. generating an apple followed by, **2**. coloring it blue. While being important, these setups fail to accurately evaluate T2I models in generating complex scenarios which require multiple iterations of reasoning. Furthermore, with the rapid development of state of the art T2I models, it is imperative to develop stringent benchmarks.

To address this gap, we develop the **AcT2I**

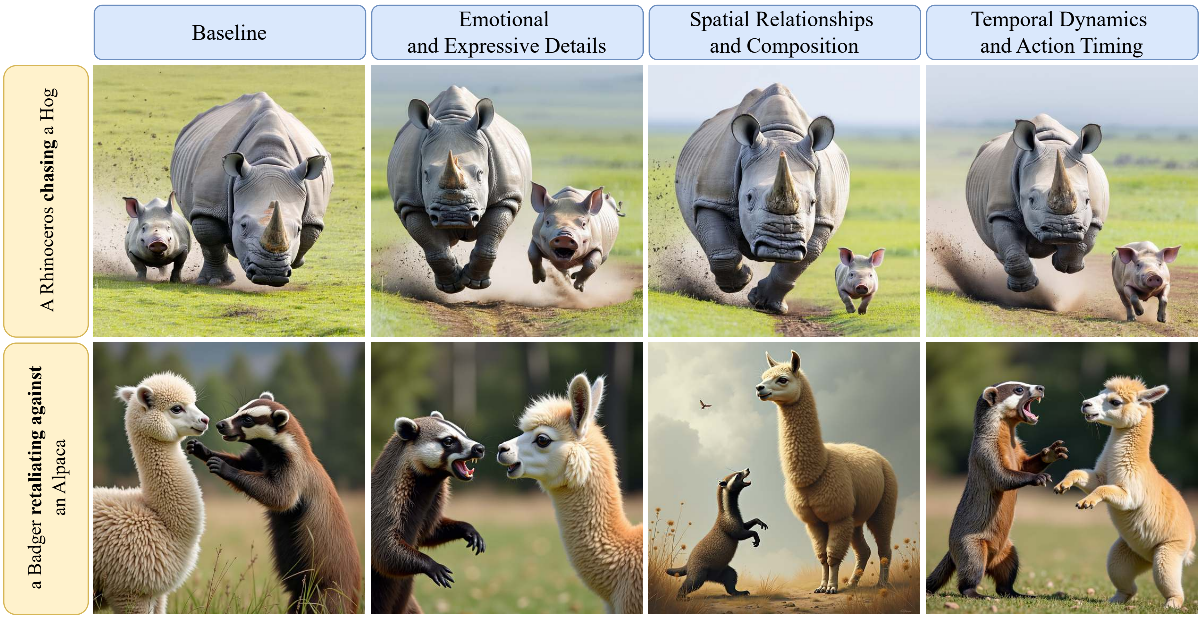

Baseline
Emotional and Expressive Details
Spatial Relationships and Composition
Temporal Dynamics and Action Timing
A Rhinoceros **chasing** a Hog
a Badger **retaliating against** an Alpaca

Figure 2: LLM-Transformed prompts unlock richer and accurate action depiction. Compared to baseline prompts that lack dense details, LLM-guided transformation into Emotional, Spatial, and Temporal dimensions generates images exhibiting more compelling action dynamics, finer expressive details, and improved subject placement accuracy. (See Appendix B for an extended qualitative analysis.)

benchmark, to evaluate Text-to-Image models in their ability to generate images of action-centric scenarios. To the best of our knowledge, action depiction has not been studied for T2I models. In this work, we aim to define this problem statement, benchmark existing models and develop baseline methods, with the ultimate goal of aligning T2I models with human interpretation of actions, which are inherently complex. For example, depicting "*a viper coiling around a duck*" involves more than just object placement and spatial composition; it requires understanding of relative object proportions, temporal dynamics and appropriate emotional expressions.

We develop the **AcT2I** evaluation benchmark by sampling a total of 20 actions from the Animal Kingdom (Ng et al., 2022) dataset, covering 100 animals, developing a total of 125 prompts. With this evaluation suite, we perform a comprehensive human evaluation, across 5 state of the art T2I models. Our key finding is that - existing T2I models struggle to generate images that accurately depict realistic actions based on textual prompts. These models tend to overfit to conventional actions associated with specific animals and fail to capture the nuanced details necessary to convey a given action effectively. Our next observation is that providing dense information in the text prompt improves performance in action generation by a significant margin. Therefore, we propose a test-time Large Language Model (LLM) (Touvron et al., 2023; OpenAI et al., 2024) guided knowledge distillation pipeline that enhances the text prompt across multiple dimensions, leading to upto *3x* gains in performance, as shown in Figure 2.

To summarize, our contributions are as follows:

- We develop the **AcT2I** benchmark to evaluate the ability of T2I models to generate images from textual prompts that describe actions. We evaluate a total of 25 actions across 100 animals, sampled from the Animal Kingdom (Ng et al., 2022) dataset.
- Our findings, based on an extensive evaluation of 5 state-of-the-art T2I models, reveal a significant limitation in their ability to generate accurate and realistic action depictions.
- We develop a training-free LLM-guided knowledge distillation technique that injects dense descriptions into prompts across 3 dimensions - spatial, temporal and emotional; and find large gains in performance, such as a 73% improvement in the performance of Stable Diffusion 3.5 Large.

2 Related Works

2.1 Text-to-Image Models

Early Text-to-Image (T2I) models focused on generating simple, often low-resolution images directly from textual prompts. This changed with Stable Diffusion (Rombach et al., 2022), which pioneered latent-space processing using VQGAN (Esser et al., 2021), enabling scalable, high-fidelity image generation. Subsequent efforts, such as Latent Consistency Models (Luo et al., 2023) and InstaFlow (Liu et al., 2023), have further optimized aspects like generation speed and image quality. Recently, models including FLUX.1-dev (Labs, 2024) and Stable Diffusion 3 (Esser et al., 2024) have pushed the boundaries of compositional accuracy and realism. However, despite these advancements, current T2I models often struggle with capturing intricate relationships and complex interactions, motivating the need for additional techniques to enhance semantic understanding.

2.2 Knowledge Distillation from Large Language Models

Descriptive captions have proven to improve image generation in text-to-image models (Betker et al., 2023). Knowledge Distillation (KD) provides a pathway to improve T2I models by transferring semantic and contextual knowledge from Large Language Models (LLMs) without additional full-scale retraining. For instance, KD-DLGAN (Cui et al., 2023) leverages generative distillation to enhance image diversity and quality even under limited data conditions. Beyond improving visual realism, KD can bridge the gap between textual semantics and visual representations, facilitating tasks like visual question answering and image captioning. Augmenting models like CLIP (Radford et al., 2021) with LLM-derived knowledge has shown promise in improving vision-language alignment (Dai et al., 2022). Nevertheless, current KD approaches often assume static relationships (Feng et al., 2024; Wu et al., 2024; Datta et al., 2023; Zhong et al., 2023) and lack the ability to handle dynamic, action-centric scenarios. This limitation underscores the need to adapt KD techniques for more sophisticated tasks where temporal and relational dynamics play a critical role.

2.3 Relational Understanding in Generative Models

While T2I models now excel at producing photorealistic images, they remain limited in their capacity to generate coherent relational scenes. Existing SOTA models—such as Stable Diffusion 3.5 Large, DALL-E 3 (Betker et al., 2023), and FLUX.1-dev—often fail to correctly depict scenarios like “a cat chasing a mouse under a table,” yielding images that lack logical spatial arrangements or contextual correctness (Chatterjee et al., 2024; Conwell and Ullman, 2022; Lian et al., 2023). The core challenge is that these models typically do not capture fine-grained relational cues, making it difficult to represent dynamic interactions or hierarchical relationships (Fu et al., 2024). Benchmarks like the Textual-Visual Logic Challenge (Xiong et al., 2025) highlight these shortcomings, focusing on compositional and logical consistency rather than isolated attributes.

In this work, we develop a benchmark that requires relational reasoning of a complex form – generating the correct action between two entities. We empirically establish that this is indeed a hard problem for existing T2I models and develop a simple baseline method to improve upon this existing shortcoming.

3 Benchmarking T2I Models on AcT2I

3.1 Experimental Setup

We evaluate 5 T2I models — Stable Diffusion 3.5 Large (Esser et al., 2024), AuraFlow[1], FLUX.1-dev (Labs, 2024), Playground v2.5 (Li et al., 2024), and PixArt-Σ (Chen et al., 2024); with these models varying across their pre-training data, diffusion architecture and text encoders. We sample 4 images per prompt to maintain consistency. All image generations were performed using publicly available model checkpoints with parameter settings, unless otherwise specified. All experiments were run on a NVIDIA A100 GPUs.

3.2 Prompt Generation

All our prompts consist of 2 entities and 1 action relationship; example prompts are enumerated in Table 1. Our entities and actions are sourced from the Animal Kingdom Dataset (Ng et al., 2022). We choose this setup because it enables to evaluate animals from diverse taxonomies (*for example*, mammals and reptiles) and sample actions of varying

[1] https://huggingface.co/fal/AuraFlow-v0.2

A	B	Action	Text
Goose	Turkey	competing for dominance with	A goose competing for dominance with a turkey
Boar	Giraffe	retaliating against	A boar retaliating against a giraffe
Weasel	Snake	fleeing from	A weasel fleeing from a snake
Duck	Duck	fighting	A duck fighting a duck
Gorilla	Dog	grooming	A gorilla grooming a duck

Table 1: Examples of text inputs from the AcT2I dataset for a pair of animals (A, B) and an action between them.

kinds. Furthermore, unlike human image generation (More details in Appendix A), T2I models do not exhibit issues such as disfigurement when generating animal images. This allows us to focus exclusively on evaluating their ability to depict actions accurately.

We cover 25 actions, generating 5 prompts/action, each instantiated with a unique animal-animal combination, leading to a total of 125 prompts. Our prompts cover both actions naturally associated with a given animal and those that are less typical, ensuring coverage of both in-distribution and out-of-distribution scenarios. Overall, our benchmark ensures a broad coverage of action types, evaluating the models' compositional reasoning, and facilitate a meaningful assessment of their ability to depict nuanced animal interactions.

3.3 Annotation Setup

A total of 25 annotators, hired via Amazon Mechanical Turk where each image was independently rated by 3 annotators. Each annotator was instructed to answer "Yes" or "No" to the question: "Does the image accurately depict the action described in the prompt?". We define the acceptance rate as the proportion of images receiving a “Yes” from a majority of annotators. This binary evaluation helps isolate whether models can convey the intended action rather than focusing on nuanced aesthetic qualities. More details are presented in Appendix E.

3.4 Benchmarking Results

Overall Performance: As shown in Figure 1, Stable Diffusion 3.5 Large achieves the highest acceptance rate (48%), followed by FLUX.1-dev (45%) and AuraFlow (44%). In contrast, PixArt-Σ and Playground v2.5 lag behind at 29% and 27%, respectively. These results indicate a considerable performance gap, with no model surpassing a 50% acceptance rate across challenging action-centric prompts, indicating that none of the models get majority of the images correct.

Category-Specific Trends: In Figure 3 we elaborate upon the acceptance rate across 2 dimensions, **1**. Animal Class vs Model Performance, and **2**. Animal Class vs Action: Figure 3(1a) breaks down acceptance rates by animal class combinations. We find that birds generally yielded the most accurate depictions, followed by mammals, while models struggled at generating images containing reptiles. FLUX.1-dev excelled in Bird-Bird prompts, reaching a 72% acceptance rate, and Stable Diffusion 3.5 Large performed best under Mammal-Mammal scenarios (52%). However, both struggled with reptile-related prompts. We find that Playground v2.5—despite its low overall acceptance—performed comparatively well on Reptile-Reptile prompts (46%), surpassing even top-performing models in this category. This suggests that some models may have niche strengths or training biases that favor certain animal classes or interactions.

These findings underscore the complexity of action-centric generation tasks. Although certain models achieve moderate success in specific domains (e.g., birds), consistently depicting complex interactions across diverse species remains a significant challenge.

3.5 Quantitative Analysis

Although T2I models have advanced considerably, our evaluations reveal persistent difficulties in accurately depicting complex, action-centric scenes. Figure 3(2a) details acceptance rates across various animal classes and actions, illustrating several recurring issues:

1. Incomplete Depictions: We find that a lot of images lack essential elements of the prompt. For instance, “coiling around” actions often produced headless snakes (Mammal-Reptile acceptance: 17.5%; Reptile-Reptile: 20.8%), and “[animal] fleeing from a cobra” frequently omits the cobra entirely. In multiple scenarios, an animal is entirely replaced by another or completely skipped, indicating that models struggle to maintain multiple distinct entities simultaneously.

2. Hybridization of Animals: Models occasionally fuse features of different species, yielding unnatural hybrids (e.g., a viper with a duck's head). For actions such as “pecking at”, "fleeing from", and "calling to" in the Bird-Mammal prompts, the low acceptance rate of 3/10 suggests difficulty differentiating species. Cross-class prompts like “a swan *pecking at* a crocodile” often produces visual

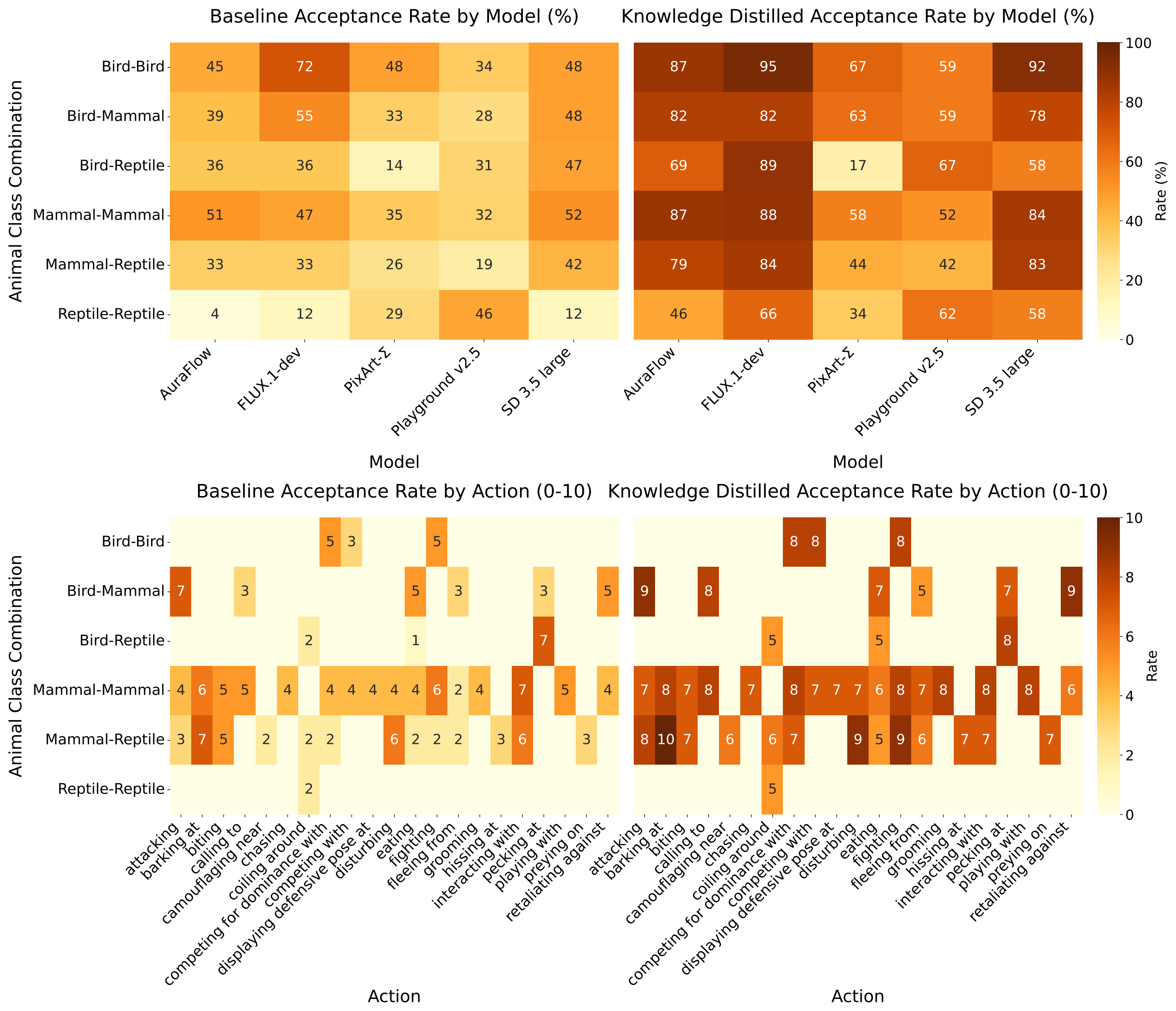


Figure 3: Heatmaps of acceptance rates for baseline and knowledge-distilled prompts. (1a) Baseline acceptance rates by model, (1b) Knowledge-distilled acceptance rates by model, (2a) Baseline acceptance rates by action, and (2b) Knowledge-distilled acceptance rates by action. Comparisons are shown across models, actions, and animal class combinations.

blends, undermining species boundaries.

3. Contextual Misrepresentation: A key challenge lies in accurately rendering the intended relationships and roles specified in a prompt. For example, cross-species interactions frequently exhibit taxonomic bias, with mammalian traits disproportionately emphasized over reptilian ones, regardless of the intended narrative. Consider, for instance, the act of "a Snake coiling around a Gecko". While the physical action of coiling may be represented, images often neglect the gecko's struggle and portrays it as willingly entangled (25% Mammal-Reptile). It is also overfitted as a negative action instead of neutral. The same taxonomic dominance is given more credence than physical dominance, with small animals able to attack big animals in a spatial area; failing to convey subtle power dynamics (43.8% Mammal-Mammal; 50.3% Bird-Bird).

4. Spatial and Positional Inaccuracy: Scenarios requiring careful scaling and depth cues are mishandled. "A bird landing on an elephant" often showed disproportionate sizes, while "perched atop a tall giraffe" lacked proper perspective. Such misalignments indicate a struggle to represent realistic spatial relationships.

5. Emotional and Expressive Inaccuracy: Prompts implying aggression or social nuance frequently produced incorrect images. For example, "retaliating against" in Bird-Mammal contexts reached only 46.7% acceptance, rarely capturing the intended hostility. Similarly, "grooming" interactions (35.6% Mammal-Mammal) lacked the gentle or intimate postures expected.

6. Temporal Dynamics and Action Timing: Actions involving movement, such as "chasing," were often rendered statically. With only 38.7% acceptance in Mammal-Mammal combinations, dynamic sequences appeared frozen in a single frame, lacking the temporal cues necessary to convey motion and directionality.

Collectively, these challenges underscore that current T2I models struggle with tasks demanding nuanced relational, spatial, emotional, and temporal understanding. Such deficiencies motivate our subsequent efforts to enrich prompts with semantic guidance to improve action depiction.

4 Improving Action Generation with LLM-guided Knowledge Distillation

To address the persistent challenges in T2I generation identified earlier, we propose a training-free prompt enrichment strategy that leverages Large Language Models (LLMs). Specifically, we use GPT-4 (OpenAI et al., 2024) to infuse additional semantic guidance into prompts. Rather than opting for retraining T2I models or modifying their architectures, we focus on enhancing the textual inputs directly. This approach is a lightweight and flexible intervention, aiming to provide clearer instructions that help models better capture relational, emotional, and temporal nuances (see Appendix D.4 for few-shot results and Appendix D.5 for the open- vs. closed-source LLM comparison).

4.1 Dimensions for Prompt Expansion

We systematically enrich prompts along three key dimensions: *spatial*, *emotional*, and *temporal*. Spatial guidance clarifies relative positioning, size, and depth; emotional cues emphasize behavioral expressions and postures; and temporal hints convey motion and sequential dynamics. By isolating these aspects, we can precisely target common failure modes in T2I generation.

4.2 Methodology

For each original prompt, we instruct GPT-4 to add semantic depth tailored to one of the three dimensions. This involves specifying the animals' relative positions, emotional states, or motion cues more explicitly. The enriched prompts thus serve as more detailed "blueprints" for the T2I model, potentially reducing ambiguity and guiding it toward more accurate renderings.

We provide illustrative templates for each dimension in Appendix D.2. These templates demonstrate how a prompt can be transformed to highlight specific spatial, emotional, or temporal aspects without fundamentally altering the underlying scenario. Importantly, these transformations are prompt-specific: each prompt's enrichment depends on its initial wording and context. By applying the template guidelines flexibly, we can adapt the semantic enrichment process to a wide range of action-centric scenes, ensuring that the resulting prompts remain coherent, contextually relevant, and aligned with the desired narrative.

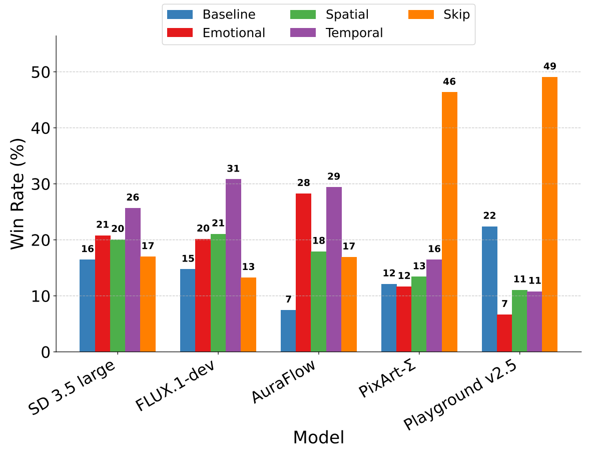


Figure 4: Aggregated user preferences (%) for image generation models across compositional dimensions. Win rates show how often users preferred a dimension over others. Skip rates represent cases where no dimensions adequately captured action-related elements.

4.3 Evaluation of Distillation Techniques

Knowledge distillation significantly enhances the performance of T2I models, particularly in capturing temporal, emotional, and spatial nuances. Figure 4 illustrates the win rate of each model across various dimensions. Stable Diffusion 3.5 Large, Flux.1-dev, and AuraFlow demonstrate notable user preference for LLM-guided enriched prompts. Temporal Distillation emerges as the most preferred option across all models, followed by Emotional and Spatial Distillation. Conversely, PixArt-Σ and Playground v2.5 exhibit limited efficacy in utilizing descriptive prompts. Subsequent paragraphs provide an in-depth analysis of each dimension's performance across different animal class pairs and actions. Our word cloud analysis (Figure 5) reveals the most frequent terms in our enriched prompt database across dimensions. Emotionally Distilled prompts feature a high frequency of expressive terms, effectively capturing subject

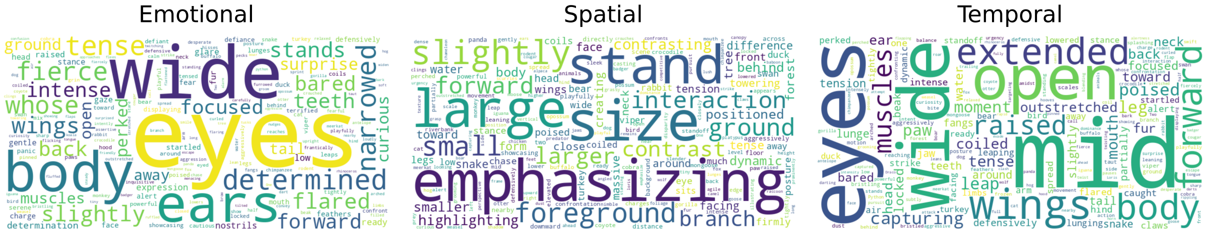


Figure 5: Word clouds summarizing key semantic elements enriched in prompts for each dimension: Emotional (left), Spatial (center), and Temporal (right).

emotions and moods. Spatially Distilled prompts emphasize precise locational descriptors, ensuring accurate spatial relationships. Temporal Distillation incorporates temporal markers, enhancing the depiction of dynamic sequences. Additionally, we conduct a POS Analysis, detailed in Appendix D.3.

Temporal Distillation: We find that temporal enrichment causes the largest improvement in performance. For example, competitive actions, which often involve nuanced sequences of dominance and retaliation, see a +274% improvement with temporal prompts. Bird-Bird interactions, known for intricate hierarchical displays, achieve a +365% gain. Specific actions like "hissing at" and "retaliating against" improve by +431% and +363%, respectively, highlighting the critical role of action timing and motion cues in disambiguating complex behaviors.

Emotional Distillation: Emotional guidance refines subtle behavioral and expressive details, substantially boosting fidelity in close-range or tension-filled scenarios. Feeding actions, which depend on accurately depicting predatory and defensive postures, benefit by +223%, while "chasing" and "coiling around" improve by +397% and +382%, respectively. In reptile-reptile interactions, emotional cues lead to a +891% improvement, underscoring how clear affective states help models represent inherently less familiar or visually subtle animal dynamics.

Spatial Distillation: Spatial enrichment ensures correct positional relationships and size contrasts. While its impact is modest in dynamic scenarios, it still provides meaningful gains for actions reliant on correct vantage points. For example, "calling to" improves by +307%, and "hissing at" sees a +292% gain under spatial prompts. These enhancements confirm that clearly specifying spatial arrangements can complement temporal and emotional cues, particularly for stationary or less overtly dynamic interactions.

Baseline and Category Dependencies: Interestingly, certain categories remain challenging, with social actions showing relatively modest gains (+37%), and baseline prompts outperforming enriched ones in some aggressive and social scenarios. Bird-Reptile interactions stand out at baseline (0.278), indicating that even without enrichment, some combinations are inherently easier.

Overall, these quantitative insights validate the qualitative claims. Temporal cues best tackle dynamic and abstract actions, emotional details help articulate close-range or expressive interactions, and spatial guidance refines positional accuracy. While no single technique solves all shortcomings, dimension-specific enrichment—particularly temporal—offers a significant step toward more nuanced, contextually accurate T2I image generation.

4.4 What about automated metrics?

We explore CLIPScore (Hessel et al., 2021) and a DINOv2(Oquab et al., 2023) based metric (DinoScore) to automatically evaluate the generated images. The objective was to assess whether automated metrics are reliable for evaluating action generation in images. CLIPScore is a reference-free evaluation metric that leverages the capabilities of the CLIP model to assess semantic alignment between textual descriptions and images. The DINOv2-based pipeline has 2 steps : **1** Extracting the most relevant frame from Animal Kingdom videos for a given action (using CLIPScore), and **2** comparing the DINOv2 features of the extracted frame and the generated image of the T2I models. As shown in Figure 6, automated metrics exhibit minimal differentiation between dimensions and fail to correlate with human evaluations on the AcT2I benchmark. Although the means in Figure 6 appear almost identical, this reflects a shared insensitivity to fine-grained action cues rather than genuine agreement; a targeted probe (Appendix C.2) shows that both metrics assign nearly the same scores to correct and mismatched captions. This discrepancy underscores the need for more sophisticated metrics that better align with human perceptual judgment. We discuss this in more depth in Appendix C.1.

Note on multimodal LLM evaluators. A pilot study with the VQA-capable model `llava-v1.6-vicuna-13b-hf` is discussed in Appendix C.3. Its limited accuracy (≈70 % overall, 62 % on semantic queries) suggests current multimodal LLMs still struggle with interpreting fine-grained action semantics.

5 Conclusion

Despite significant advancements in Text-to-Image (T2I) synthesis, current models exhibit limitations in accurately representing nuanced actions, highlighting a gap between model capabilities and real-world expectations. Our evaluation reveals that state-of-the-art models achieve only a 48% acceptance rate, underscoring the difficulty in capturing the implicit visual cues crucial for representing dynamic scenarios.

To address this challenge, we introduced knowledge-distilled techniques targeting three key dimensions: temporal dynamics, emotional expressiveness, and spatial relationships. Temporal distillation emerged as the most impactful, significantly enhancing the depiction of dynamic actions. Emotional and spatial distillations complemented this by refining subtle behavioral and positional elements, respectively.

While automated metrics like CLIPScore and DinoScore offer valuable insights, they fall short in capturing the nuanced improvements achieved through our techniques. Human evaluations remain the gold standard for assessing semantic fidelity and realism in complex T2I outputs.

Future research should focus on enhancing relational and temporal grounding in vision-language (VL) models to better capture the implicit visual cues critical for nuanced action representation. Furthermore, the development of robust automated metrics capable of accurately evaluating complex T2I outputs remains a crucial area for future exploration, ensuring that progress in this field can be effectively measured and validated.

6 Limitations

While this study demonstrates the efficacy of training-free semantic enrichment, two primary limitations merit consideration. First, the observed performance gains are contingent upon the underlying Text-to-Image (T2I) model's capacity to effectively leverage densely enriched prompts. The extent of this capacity may vary depending on the initial prompt complexity and the quality of the Large Language Model (LLM)-generated output. Second, the evaluation methodology relies heavily on human assessment due to the inherent limitations of current automated metrics, which are often inadequate for capturing subtle semantic nuances comprehensively. While our findings indicate significant improvements in contextual understanding for T2I models, they also raise potential societal concerns, including the potential misuse of more realistic imagery and the propagation of inherent biases present within the training data. Future research must address these ethical considerations to ensure responsible applications of these techniques and mitigate potential negative consequences.

References

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf, 2(3):8.

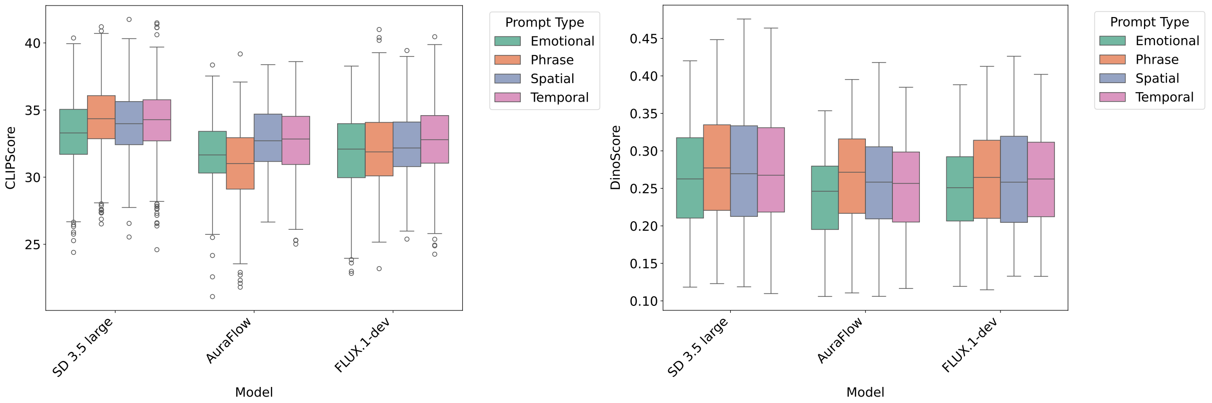
Agneet Chatterjee, Gabriela Ben Melech Stan, Estelle Aflalo, Sayak Paul, Dhruba Ghosh, Tejas Gokhale, Ludwig Schmidt, Hannaneh Hajishirzi, Vasudev Lal, Chitta Baral, et al. 2024. Getting it right: Improving spatial consistency in text-to-image models. In European Conference on Computer Vision, pages 204–222. Springer.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. 2024. Pixart-σ $: Weak - to - strongtrainingofdiffusiontransformerfor4ktext - to - imagegeneration.$ Preprint, arXiv:2403.04692.

Colin Conwell and Tomer Ullman. 2022. Testing relational understanding in text-guided image generation. arXiv preprint arXiv:2208.00005.

Kaiwen Cui, Yingchen Yu, Fangneng Zhan, Shengcai Liao, Shijian Lu, and Eric P Xing. 2023. Kd-dlgan: Data limited image generation via knowledge distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3872–3882.

Wenliang Dai, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. 2022. Enabling multimodal generation on clip via vision-language knowledge distillation. arXiv preprint arXiv:2203.06386.



(a) CLIPScore (higher is better)

(b) DinoScore (higher is better)

Figure 6: Performance of Automated Evaluation Metrics on the AcT2I Benchmark. Neither metric aligns well with human annotator preferences.

Siddhartha Datta, Alexander Ku, Deepak Ramachandran, and Peter Anderson. 2023. Prompt expansion for adaptive text-to-image generation. arXiv preprint arXiv:2312.16720.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In Forty-first International Conference on Machine Learning.

Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 12873–12883.

Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2024. Layoutgpt: Compositional visual planning and generation with large language models. Advances in Neural Information Processing Systems, 36.

Xingyu Fu, Muyu He, Yujie Lu, William Yang Wang, and Dan Roth. 2024. Commonsense-t2i challenge: Can text-to-image generation models understand commonsense? arXiv preprint arXiv:2406.07546.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. 2024. Geneval: An object-focused framework for evaluating text-to-image alignment. Advances in Neural Information Processing Systems, 36.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. Advances in Neural Information Processing Systems, 36:78723–78747.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 4401–4410.

Black Forest Labs. 2024. Flux. https://github.com/black-forest-labs/flux.

Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. 2024. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation. Preprint, arXiv:2402.17245.

Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. 2023. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. arXiv preprint arXiv:2305.13655.

Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, et al. 2023. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In The Twelfth International Conference on Learning Representations.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. 2023. Latent consistency models: Synthesizing high-resolution images with few-step inference. arXiv preprint arXiv:2310.04378.

Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. 2022. Animal kingdom: A large and diverse dataset for animal behavior understanding. Preprint, arXiv:2204.08129.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. Preprint, arXiv:2303.08774.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PMLR.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 10684–10695.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Tsung-Han Wu, Long Lian, Joseph E Gonzalez, Boyi Li, and Trevor Darrell. 2024. Self-correcting llm-controlled diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6327–6336.

Peixi Xiong, Michael Kozuch, and Nilesh Jain. 2025. Textual-visual logic challenge: Understanding and reasoning in text-to-image generation. In Computer Vision – ECCV 2024, pages 318–334, Cham. Springer Nature Switzerland.

Shanshan Zhong, Zhongzhan Huang, Weushao Wen, Jinghui Qin, and Liang Lin. 2023. Sur-adapter: Enhancing text-to-image pre-trained diffusion models with large language models. In Proceedings of the 31st ACM International Conference on Multimedia, pages 567–578.

Appendix

A Addressing Generalization Limitations

To evaluate action depiction, we aimed to assess T2I models' ability to generate images where all subjects were accurately represented. Additionally, our choice of subjects prioritized those with high affordance, capable of performing a diverse set of actions and assuming unique roles with distinct mappings. However, the current capabilities of T2I models, when tasked to generate all subjects correctly, introduce more issues. We found that generating images of humans remains a challenge, with common errors including incomplete depictions and subject disfigurement. These errors directly impact evaluation because we can only thoroughly critique semantic inaccuracies after the subjects themselves are rendered correctly. We share examples of such errors in Action depiction with two Human subjects using Stable Diffusion 3.5 Large in Figure 7a. We additionally also share the generalization capabilities of our knowledge distillation techniques to improve upon semantic details represented through each action across dimensions (spatial, emotional, and temporal) in Figure 7b. Likewise, prompts involving abstract concepts are problematic because their many visual interpretations make objective evaluation much harder, leading us to focus on more grounded scenarios. Due to their inherent characteristics, animals provide an ideal test case for evaluating the generalization performance of T2I models, making them our primary choice for this study.

B Qualitative Evidence of Semantic Enrichment

Figure 8 presents eight representative generations produced with our LLM-guided prompt enrichment. The images are grouped into two categories:

Semantically Enriched & Visually Complete (left block, four examples) — spatial layouts, emotional expressions, and temporal framing all align with the target action, yielding high-fidelity results (e.g., “a sheep attacking a boar”).

Semantically Enriched but Visually Flawed (right block, four examples) — the same semantic cues are present, yet the base T2I model introduces pixel-level errors such as missing subjects or anatomical distortions (e.g., “a weasel fleeing from a bear”).

Model Name	CLIPScore	DinoScore
Auraflow	0.30	0.20
FLUX.1-dev	0.38	0.30
PixArt-Σ	0.38	0.24
Playground v2.5	0.56	0.35
SD 3.5 large	0.38	0.24

Table 2: Alignment evaluation of Automated metrics with Human preferences, where 1 indicates full alignment, and 0 indicates no alignment.

Across both groups, three high-level dimensions are consistently evident:

- **Spatial relationships**: correct relative placement and orientation of agents.
- **Emotional cues**: facial expressions or body posture that match the action context.
- **Temporal framing**: a frame that captures the peak moment of the action.

The persistence of these cues in visually flawed outputs indicates that semantic enrichment operates independently of pixel-level rendering quality, complementing the quantitative gains reported in Section 4.

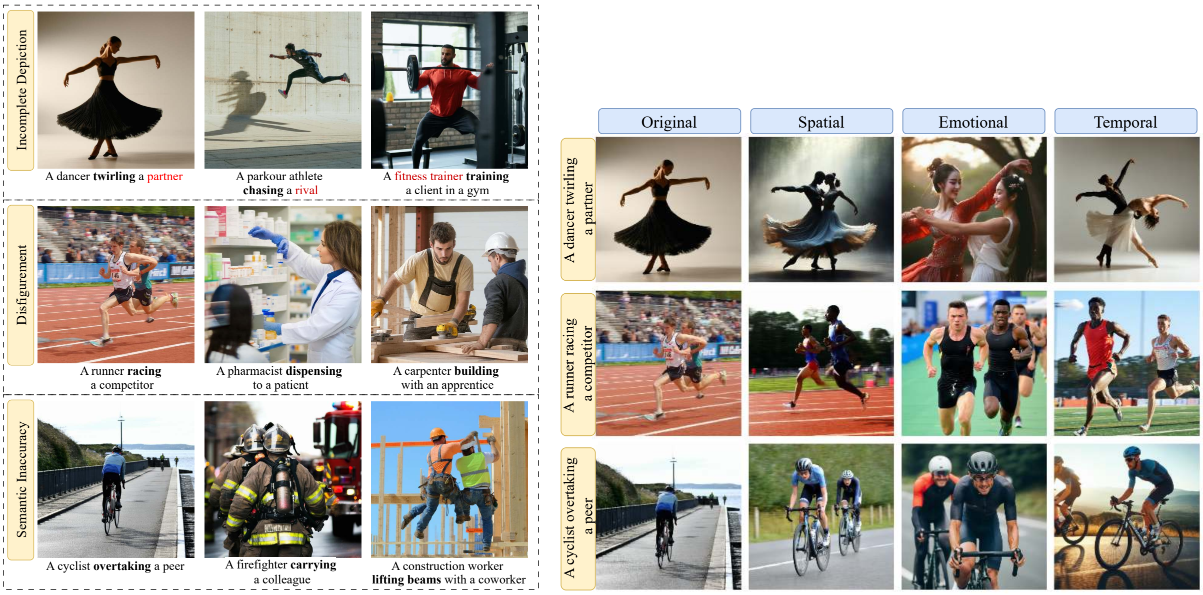
C More results on Automated Metric

C.1 DinoScore Evaluation

We collect human annotation preferences and derive a consensus from the three reviews for each sample. The preferred dimension is then compared to the highest CLIPScore and DinoScore, respectively. Table 2 presents the alignment of these two automated evaluation metrics with human preferences. Given four possible choices, the baseline alignment is 0.25. The highest alignment observed among the top three models is 0.38 for CLIPScore and 0.30 for DinoScore. These results indicate that while there is some overlap, the metrics exhibit significant limitations in capturing subtle semantic improvements.

C.2 Action-specific probe for metric granularity

We generated an image with Stable Diffusion 3.5 for the prompt “a cat chasing a mouse”. Table 3 lists CLIPScore and DinoScore for three candidate captions— *chasing* (correct), *observing from afar*,


Incomplete Depiction
A dancer twirling a partner
A parkour athlete chasing a rival
A fitness trainer training a client in a gym
Disfigurement
A runner racing a competitor
A pharmacist dispensing to a patient
A carpenter building with an apprentice
Semantic Inaccuracy
A cyclist overtaking a peer
A firefighter carrying a colleague
A construction worker lifting beams with a coworker
Original
Spatial
Emotional
Temporal
A dancer twirling a partner
A runner racing a competitor
A cyclist overtaking a peer

(a) Examples of generation failures for human actions. (b) Semantic improvements achieved through knowledge distillation.

Figure 7: Challenges in generating images of human actions with Stable Diffusion v3.5 Large. Panel (A) highlights common errors— incomplete depictions (missing subjects or objects), disfigurement (physical anomalies), and semantic inaccuracies (misrepresented actions). Despite these errors, panel (B) demonstrates the generalization capabilities of our technique on human samples.

Candidate caption	CLIP↑	DINO↑
Cat **chasing** a mouse	26.9	0.257
Cat **observing** from afar	25.9	0.286
Cat **attacking** a mouse	25.6	0.290

Table 3: CLIPScore and DinoScore for the probe image (Figure 9). The small gaps between correct and mismatched captions illustrate each metric's coarse granularity.

and *attacking*. The correct caption scores highest in both metrics, yet the margin over incorrect captions is small (< 5 % for CLIPScore, < 0.04 absolute for DinoScore), confirming that the metrics capture high-level entity alignment but struggle with nuanced action semantics.

C.3 Multimodal LLM Evaluation

We also explored using the multimodal LLM `llava-v1.6-vicuna-13b-hf` as a VQA-style evaluator. For each prompt, we automatically generated ten tailored questions covering salient scene attributes. On a proof-of-concept (POC) set of eight prompts, LLaVA's answers matched human annotations 70% of the time; accuracy dropped to 62% on questions probing nuanced semantic details (e.g. object–action relations). Given this modest performance and the cost of querying LLMs, we did not pursue this avenue further at scale.

D Prompt based Analysis

D.1 Action Template Taxonomy

In Table 4, we share the action templates used to generate prompts, categorized by their plausibility tiers (Highly Plausible, Moderately Plausible, and Less Plausible). These templates guided the selection of animals and actions to ensure a broad range of complexity and contextual requirements.

D.2 Prompt Distillation Guidelines

Below are the guidelines we used to enrich prompts with spatial, temporal, and emotional details with an average token count of 47, 42, and 46, respectively. These were applied using an LLM (GPT-4o) to create enriched versions of the original prompts, providing more explicit cues that aid T2I models in generating contextually accurate images. Through prompts, LLM was instructed to enhance prompts for text-to-image tasks through knowledge distillation in [dimension], followed by an explanation of what was expected. Each instruction concluded with a message of keeping the enhanced prompt concise yet detailed and aiming for approximately 50-70 tokens while prioritizing clarity over length.

1. Spatial Relationships and Composition:

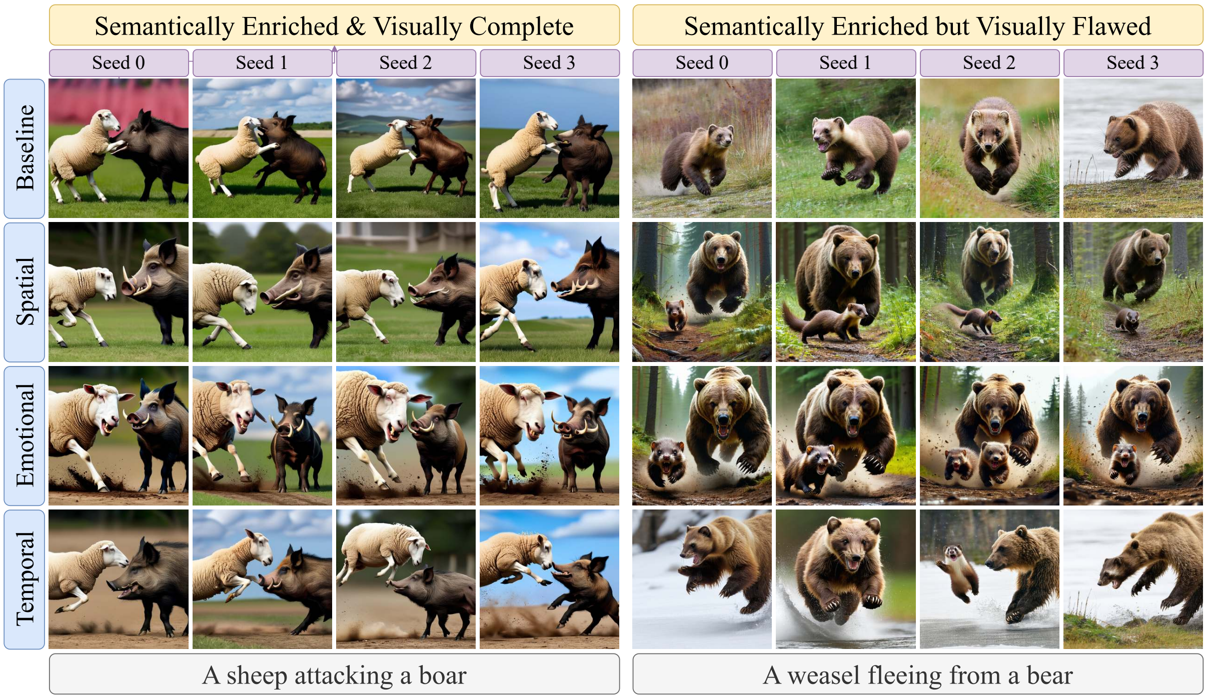

Semantically Enriched & Visually Complete
Semantically Enriched but Visually Flawed
Seed 0
Seed 1
Seed 2
Seed 3
Seed 0
Seed 1
Seed 2
Seed 3
Baseline
Spatial
Emotional
Temporal
A sheep attacking a boar
A weasel fleeing from a bear

Figure 8: Qualitative grid illustrating our LLM-guided prompt enrichment. **Left block (“Semantically Enriched & Visually Complete”)**: spatial layout, emotional cues, and temporal framing all align with the intended action. **Right block (“Semantically Enriched but Visually Flawed”)**: the same semantic cues are present, but pixel-level errors such as missing subjects or distortions remain, highlighting that enrichment injects reliable semantics even when rendering fidelity fails.

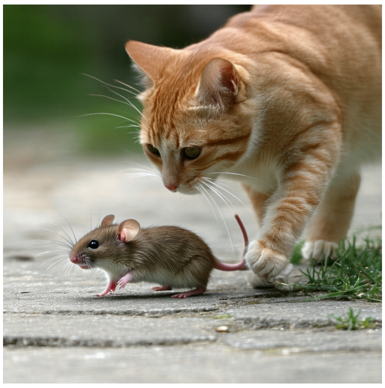


Figure 9: Stable Diffusion 3.5 image generated for “cat chasing a mouse.”

> Make implicit spatial details explicit to improve the prompt, while keeping it concise and focused. Pay attention to:
> **Positional Accuracy**: Clearly specify how animals are positioned relative to each other.
> **Depth and Perspective**: Indicate scaling and perspective for appropriate distance and interaction.
> **Example**: Instead of "a bird lands on an elephant", say "a small bird gently lands atop a towering elephant’s back, highlighting their size difference".

2. Temporal Dynamics and Action Timing:

> Make implicit temporal and action details explicit to improve the prompt, while keeping it concise and focused. Emphasize:
> **Optimal Freeze-Frame Selection**: Capture the most expressive moment of the action.
> **Motion Representation**: Use visual cues like dynamic posture to imply movement.
> **Example**: Instead of "a cheetah chases a gazelle", say "a cheetah mid-stride with muscles tensed, closely pursuing a gazelle in full sprint".

3. Emotional and Expressive Details:

> Make implicit emotional details explicit to improve the prompt, while keeping it concise and focused. Include:
> **Facial Expressions**: Depict emotions appropriate to the action.
> **Body Language**: Use posture and movement to enhance emotional portrayal.
> **Example**: Instead of "a puppy chases a kitten", say "a playful puppy with a wagging tail chases a kitten that’s glancing back with a mischievous grin".

D.3 POS Tag Analysis

Through POS tagging, we analyzed the prompts to interpret the prompt distillation. We discovered that all the enriched prompts reduced the usage of proper nouns by 70%-80% while verbs and nouns increased by 5x-6x and 10x-15x, respectively, across all dimensions. Intra-dimensional analysis revealed that adjectives were 1.5x-2x more frequent in the Emotional dimension compared to

<table>
<tr><th>Plausibility Level</th><th>Action Templates</th></tr>
<tr><td>Highly Plausible Actions</td><td>[reptile|mammal|bird] attacking [reptile|mammal|bird]
[mammal] chasing [mammal|bird|reptile]
[mammal|reptile] eating [mammal|bird|reptile]
[mammal|reptile|bird] fleeing from [mammal|bird|reptile]
[reptile|mammal|bird] competing with [reptile|mammal|bird]
[mammal] fighting [mammal|reptile]
[bird] fighting [bird]
[mammal|reptile|bird] disturbing [reptile|mammal|bird]
[reptile|mammal|bird] biting [reptile|mammal|bird]
[mammal] playing with [mammal]
[bird] competing for dominance with [bird]
[mammal|bird] grooming [mammal|bird]
[mammal] retaliating against [reptile|mammal|bird]</td></tr>
<tr><td>Moderately Plausible Actions</td><td>[mammal] barking at [mammal|reptile]
[reptile] hissing at [mammal|bird]
[reptile|mammal] competing for dominance with [reptile|mammal]
[reptile] coiling around [mammal|bird|reptile]
[reptile] preying on [mammal|bird]
[bird|mammal] calling to [bird|mammal]
[bird|mammal] fleeing from [mammal|bird]
[reptile] camouflaging near [mammal|bird]</td></tr>
<tr><td>Less Plausible Actions</td><td>[bird] pecking at [reptile|mammal]
[mammal|bird] fleeing from [reptile]
[reptile|mammal|bird] interacting with [mammal|bird|reptile]
[reptile|mammal] displaying defensive pose at [reptile|mammal|bird]</td></tr>
</table>

Table 4: Action templates grouped by plausibility. These templates guided prompt creation, ensuring diverse scenarios from simple to highly complex and context-dependent.

other dimensions, while the Spatial dimension exhibited significantly higher usage of determiners, adposition, adverbs, and pronoun-particles. Overall, the top 10 most frequently used adjectives and verbs across each dimension were found to be in alignment with the intended meaning of each.

D.4 Few Shot Experiments

We conducted a preliminary experiment employing the few-shot prompting technique using GPT 4o and Gemini 2.0 Flash, utilizing three instances of original enriched prompts. The results of a blind review comparison between images generated from original enriched prompts and few-shot outputs of both models indicated comparable performance.

D.5 Open Source vs Closed Source LLMs

The cost of LLM APIs remains a key concern for the practical utility of our technique. To address this, we conducted a small-scale analysis comparing the closed-source model GPT-4o with the open-source `meta-llama/Llama-3.3-70B-Instruct`. A blind review reveals that both models perform comparably, thereby alleviating cost-related concerns.

D.6 Additional Analysis

Figure 10 shows a diverged bar graph comparing baseline prompts versus dimension-enriched prompts across various action categories. This visualization illustrates how each enrichment dimension shifts performance relative to the baseline.

E Annotation Details

Annotator Instructions: 3 independent annotators evaluated each generated image by answering: "Does the image truly represent the action in the prompt?" Annotators considered correctness of the entities, plausibility of the depicted action, and sub-

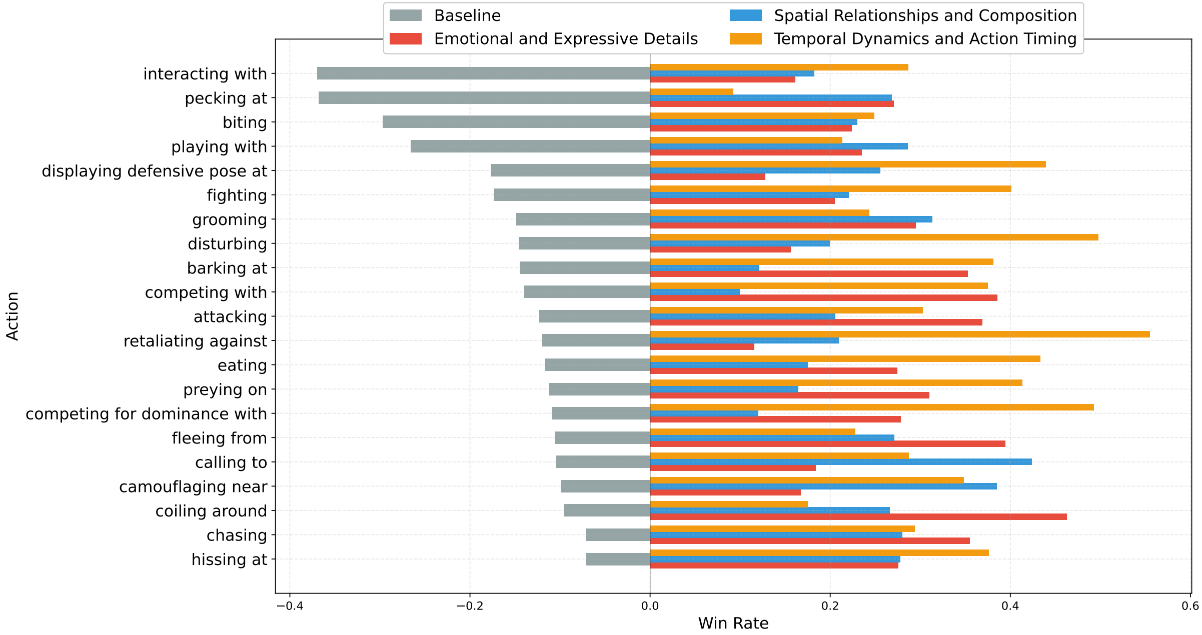


Figure 10: Diverged bar graph comparing win rates of baseline and dimension-enriched prompts across different action categories.

tle cues like emotions, spatial arrangement, and implied motion. They were encouraged to look beyond surface-level accuracy and assess whether the scene convincingly captured the intended relationships and dynamics.

Annotator Details: We crowdsourced on Amazon Mechanical Turk, 25 annotators in total completed the blind review.

Privacy and Ethics: Our dataset involves animal subjects with no personal data. The Animal Kingdom dataset and generated images are free of sensitive human information, ensuring compliance with ethical research guidelines and no privacy concerns.

F Implementation Details

We used publicly available model checkpoints and default parameters for image generation. Each prompt was rendered with four random seeds per model. Hyperparameters such as guidance scale, sampling steps, and resolution were kept consistent across models and conditions.

For enrichment, we employed GPT-4 with fixed temperature and token limits to ensure consistent output quality. Minor adjustments were made to each enriched prompt until it provided clear semantic guidance without altering the core meaning of the original prompt.