Predicated Diffusion: Predicate Logic-Based Attention Guidance for Text-to-Image Diffusion Models

Supplementary Material

A. Detailed Experimental Setting

A.1. Implementation Details

Guidance Calculation We confirmed that Predicated Diffusion that directly employed the loss functions in Section 3 worked well. Nonetheless, to further improve the quality of generated images, we implemented several modifications.

The loss functions are defined using either the summation or product over pixels i, which increase or decrease unboundedly as the image is scaled. To make these loss functions more suitable for neural networks and to prevent excessively large values, we replaced the summation with the arithmetic mean. Additionally, we replaced the finite product in (1) with the geometric mean. The equation $P(x) \wedge P(x) = P(x)$ holds for the classical Boolean logic and Gödel fuzzy logic but does not for the product fuzzy logic. Due to our definition of the strong conjunction $\wedge$ and the assumption of $P(x) \in [0, 1]$, the product fuzzy logic leads to the inequality $P(x) \wedge P(x) \leq P(x)$. This is the reason why the loss functions can increase or decrease unboundedly.

Attention maps are often obtained by performing a softmax operation in the channel direction on the feature maps of a CNN, where each channel is linked to a single word or token. For calculating the loss function (1) for existence, we excluded the start-of-text token. Because this token is linked to the entire text, its omission ensures capturing the response to each individual word, consistent with the implementation of Attend-and-Excite. For calculating the loss functions (2), (5), and (6) using (bi)implication, we normalized the intensity of each attention map to a range of 0-1 using the maximum and minimum values. This allows us to focus on the relative positions rather than the absolute intensity (that is, existence).

Following Attend-and-Excite [2] and SynGen [30], the gradient of loss function, $\nabla_{x_t}\mathcal{L}[R]$, was multiplied by 20 for updating images. Drawing inspiration from Attend-and-Excite [2], we performed the iterative refinement at $t = T$ (that is, at the beginning of the reverse process). Specifically, before executing the very first step of the reverse process, we updated the image x under generation as $x \leftarrow x - \nabla_x\mathcal{L}[R]$ five times. Note that Attend-and-Excite performs the iterative refinement at the 10th and 20th steps until the value of the loss function falls below a certain threshold. However, we found that the very first step is crucial for modifying the layout of the generated image. This may be because the reverse process of the diffusion model is similar to gradient descent and may converge to poor local minima given poor initial values; the iterative refinement at $t = T$ helps adjust the initial values for better outcomes. The results of the ablation study are provided in Table A1, which shows that the iterative refinement slightly improved image fidelity and quality measured as similarities and CLIP-IQA.

In preliminary experiments, we found that the fidelity of the generated images improved more effectively as the number of iterative refinements at $t = T$ was increased. No degradation in quality or diversity was observed even after several dozen iterations. However, due to the limitations of available computational resources, we did not adjust for the optimal number of refinements; this is a subject for future research.

Computational Cost In practice, diffusion models are frequently combined with classifier-free guidance and negative prompts. Then, the integration of Predicated Diffusion leads to a 33 % increase in computational cost.

The original diffusion model calculates an image update conditioned on a text prompt. The classifier-free guidance uses the same diffusion model without any conditions and emphasizes the difference between conditional and unconditional updates. A negative prompt computes an update using the same diffusion model conditioned on an additional text and subtracts this update from the original one, negating the text. Consequently, the diffusion model with the classifier-free guidance and a negative prompt requires three times the computational cost of the original model.

Predicated Diffusion reuses attention maps generated during the conditional update, avoiding extra computational cost. The computational cost of defining the loss function with attention maps is negligible compared to CNNs. However, computing the gradient of this loss function, $\nabla_{x_t}\mathcal{L}[R]$, via automatic differentiation incurs a computational cost comparable to the forward computation, equivalent to additional use of the diffusion model. Therefore, when Predicated Diffusion is employed with classifier-free guidance and negative prompts, the total computational cost is equivalent to four times that of the original diffusion model, with Predicated Diffusion contributing to a 33 % increase in the overall computation.

Table A1. Ablation Study of Refinement.

	Experiment (i) for Concurrent Existence		Experiment (ii) for One-to-One Correspondence		Experiment (iii) for Possession	
Methods	**Similarity**‡	**CLIP-IQA**	**Similarity**‡	**CLIP-IQA**	**Similarity**‡	**CLIP-IQA**
Without refinement	0.348 / 0.822	0.771	0.376 / 0.801	0.764	0.339 / 0.854	0.764
With refinement	0.348 / **0.825**	**0.775**	**0.379 / 0.811**	**0.769**	**0.345 / 0.855**	**0.765**

‡Text-image similarity and text-text similarity.

A.2. Instructions to Evaluators

Three, three, and two evaluators joined Experiments (i), (ii), and (iii), respectively. They are university students aged between 19 and 21, with no background in machine learning or computer vision. The experiments were conducted in a double-blind manner: the images were presented in a random order, and the evaluators and authors were unaware of which image was generated by which model. Each image was assessed by a single evaluator to maximize the number of assessed images.

Experiment (i): Concurrent Existence We prepared 400 random prompts, each mentioning "[Object A] and [Object B]" with indefinite articles as needed, and generated 400 sets of images.

(a) By showing one image at a time at random, we asked, "Are both specified objects generated in the image?" The evaluators answered this question with one of the following options:
  1) "No object is generated."
  2) "Only one of two objects is generated."
  3) "Two objects are generated, but they are mixed together to form one object."
  4) "Two objects are generated."
  Responses 1) and 2) were categorized as "missing objects", and response 3) was categorized as "object mixture". We tallied the number of responses 1) and 2) under the lenient criterion and that of responses 1)–3) under the strict criterion.

(b) By showing a set of images generated with different methods, we asked, "Which image is the most faithful to the prompt?" The evaluators were instructed to select only one image in principle, but were allowed to select more than one image if their fidelities were competitive, or not to select any image if none were faithful.

Experiment (ii): One-to-One Correspondence We prepared 400 random prompts, each mentioning "[Adjective A] [Object A] and [Adjective B] [Object B]" with indefinite articles as needed.

(a) The same as above.

(c) At the same time as question (a), we asked, "Does each adjective exclusively modify the intended object?" The evaluators answered this question with a "Yes" or "No", and the response "No" was categorized as "attribute leakage". If one or both of the two specified objects were not generated, that is, the response to question (a) was not 4), then question (c) became irrelevant, thereby automatically marking "No" as the response.

(b) The same as above.

Experiment (iii): Possession We prepared 10 prompts, each mentioning "[Subject A] is [Verb C]-ing [Object B]" with indefinite articles as needed. We generated 20 images for each of these prompts.

(a) The same as above.

(d) At the same time as question (a), we asked, "Is the [Subject A] performing [Verb C] with the [Object B]?" The evaluators answered this question with a "Yes" or "No", and the response "No" was categorized as "possession failure." If the response to question (a) was not 4), the response to question (d) was automatically assigned "No".

(b) The same as above.

A.3. Prompts

For Experiments (i) and (ii), we randomly selected objects and adjectives from Table A2, roughly following the experiments conducted by Chefer et al. [2]. However, we excluded "mouse" and "backpack" from the list of objects, and "orange" from the list of adjectives. The term "mouse" often led to ambiguity, as it could refer to either the animal or a computer peripheral. The term "orange" also created confusion in all methods, as it could indicate either the color or the fruit. Furthermore, it is challenging to distinguish "backpack" visually from other types of bags. To ensure the accuracy of the evaluation, we removed these terms from the lists.

We used prompts in Table A3 for Experiment (iii).

B. Additional Experiments and Results

B.1. Additional Analysis

Visualization of Attention Maps We visualized the attention maps linked to the words of interest in a given

Table A2. Candidate Words for Generating Prompts in Experiments (i) and (ii)

Object	cat, dog, bird, bear, lion, horse, elephant, monkey, frog, turtle, rabbit, glasses, crown, suitcase, chair, balloon, bow, car, bowl, bench, clock, apple
Adjective	red, yellow, green, blue, purple, pink, brown, gray, black, white

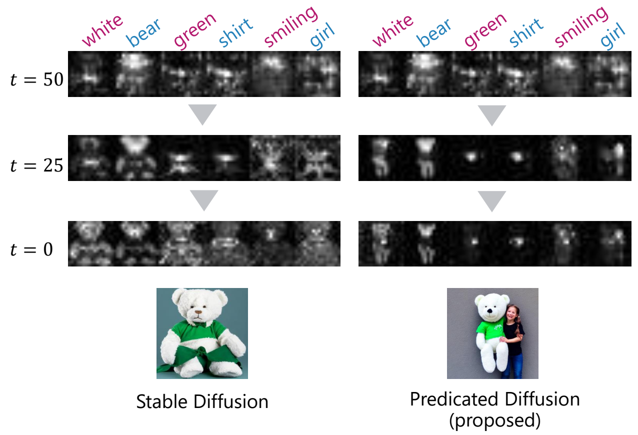


Figure A1. An example visualization of attention maps during the generation. The text prompt was "A white teddy bear with a green shirt and a smiling girl."

prompt in Fig. A1. The attention map is obtained by applying a softmax operation along the token (word) axis of the feature map of the CNN. Consequently, at most, one attention map can respond strongly at a specific pixel location. At the start of the reverse process ($t = 50$), each attention map responded to the image x to some extent. As depicted in the left half of Fig. A1, the vanilla Stable Diffusion largely maintained the intensity distributions of the attention maps until the end of the reverse process, $t = 0$. This implies that the layout of the generated image is largely influenced by random initialization, which may not accurately capture the intended meaning of the prompt. The attention map for "bear" dominates the entire space at $t = 50$ and leaves no room for the attention map for "girl" to respond, resulting in the failure to generate a girl in the image. In contrast, given the proposition $\exists x.\, Girl(x)$, Predicated Diffusion encourages the attention map for "girl" to respond, even if it means dampening the response of the attention map for "bear", thereby ensuring the presence of a girl.

Detailed Results of Experiment (iii) For reference, we summarize the actual numbers out of 20 responses for each prompt in the right half of Table A3. The successful rate varies greatly depending on the prompt, but the proposed method, Predicated Diffusion, shows the best results in almost all combinations of metrics and prompts. The sole exception is the prompt "A man holding a rabbit," where Stable Diffusion already produced satisfactory results but Predicated Diffusion deteriorated the scores. When the backbone, Stable Diffusion, can generate images faithful to the prompt, the additional guidance might disturb the generation process.

Additional Visualizations We summarize the additional visual examples of Experiments (i)–(iv) in Figs. A2–A6. While a quantitative comparison is difficult, Predicated Diffusion often retains the original layout when Stable Diffusion produces satisfactory results. This is because, as long as logical operations like implications are satisfied, Predicated Diffusion does not trigger any further changes.

B.2. Extensions to Other Logical Statements

Multi-color by Logical Disjunction Consider cases where multiple colors are specified for a single object. For instance, the prompt "a green and grey bird" implies that every part of the bird is either green or grey, not both. This statement can be represented using the disjunction by $\forall x.\, Bird(x) \rightarrow Green(x) \lor Grey(x)$. The corresponding loss function is:

$$\begin{aligned}\mathcal{L}[\forall x.\, Bird(x) &\rightarrow Green(x) \lor Grey(x)] \\ &= -\textstyle\sum_i \log(1 - A_{Bird}[i] \times (1 - A_{Green}[i]) \\ &\qquad\qquad \times (1 - A_{Grey}[i])). \end{aligned} \tag{A1}$$

When another object is introduced, one can replace the implication with a biimplication, as is the case with one-to-one correspondence.

The right six columns of Fig. A5 show example results. SynGen produced birds with a mixed hue because it was designed to equalize the intensity distributions (that is, the regions) of both specified colors and that of the bird. Conversely, Predicated Diffusion, based on predicate logic, can generate birds in the given combination of colors.

Note that, when multiple adjectives modify the same noun independently, they can be represented using the logical conjunction rather than the logical disjunction. For instance, the prompt "long, black hair" can be decomposed into two statements that can hold simultaneously: "The hair is long," and "The hair is black." Then, the prompt is represented by the conjunction of two propositions that represent these statements.

Negation by Logical Negation In the right three columns of Fig. A6, we explored the negation of a concept. Sometimes, we might wish for certain concepts to be absent or negated in the generated images. If given the prompt "a polar bear," the output will typically be an image of a polar bear depicted with snowy landscapes because of their

Table A3. Prompts for Experiment (iii) and Results for Each Prompt

[Subject A]	**[Verb C]-ing**	**[Object B]**	**Methods**	∗1	∗2	∗3	∗4
rabbit	having	phone	Stable Diffusion	12	12	15	5
			Attend-and-Excite	2	2	12	5
			Predicated Diffusion	1	1	10	9
bear	having	apple	Stable Diffusion	2	2	5	10
			Attend-and-Excite	0	1	10	1
			Predicated Diffusion	0	1	5	10
monkey	having	bag	Stable Diffusion	12	12	12	2
			Attend-and-Excite	3	6	6	6
			Predicated Diffusion	0	1	1	12
panda	having	suitcase	Stable Diffusion	13	16	19	1
			Attend-and-Excite	1	7	11	3
			Predicated Diffusion	1	2	8	7
lion	wearing	crown	Stable Diffusion	12	12	12	8
			Attend-and-Excite	0	0	0	20
			Predicated Diffusion	0	0	0	20
frog	wearing	hat	Stable Diffusion	9	10	10	3
			Attend-and-Excite	2	2	5	7
			Predicated Diffusion	1	1	4	11
man	holding	rabbit	Stable Diffusion	0	0	2	12
			Attend-and-Excite	3	4	12	4
			Predicated Diffusion	4	5	6	7
woman	holding	dog	Stable Diffusion	1	4	5	12
			Attend-and-Excite	3	8	16	2
			Predicated Diffusion	1	3	3	10
boy	grasping	soccerball	Stable Diffusion	1	1	16	4
			Attend-and-Excite	1	3	18	2
			Predicated Diffusion	0	0	14	6
girl	holding	suitcase	Stable Diffusion	1	3	9	10
			Attend-and-Excite	0	1	13	5
			Predicated Diffusion	0	0	8	12

∗1 Missing objects in the lenient criterion, ∗2 Missing objects in the strict criterion,
∗3 Possession failure, ∗4 Fidelity.

high co-occurrence rate in the dataset. One can give the prompt "a polar bear without snow," but Stable Diffusion often struggles to remove the snow, as depicted in the top row. Alternatively, we could provide a negative prompt "snow" as proposed as part of Composable Diffusion [18]. We also examined Perp-Neg, which ensures a negative prompt not to interfere with a regular prompt by projecting the former's update to be orthogonal to the latter's update [1]; it often failed to remove the snow. We consider an alternative way using predicate logic. The absence of snow is represented by the proposition $\neg(\exists x.\, Snow(x)) = \forall x.\, \neg Snow(x)$, leading the loss function:

$$\mathcal{L}[\neg(\exists x.\, Snow(x))] = -\textstyle\sum_i \log(1 - \bar{A}_{Snow}[i]), \tag{A2}$$

where $\bar{A}_w$ represents the attention map corresponding to a word w in an auxiliary prompt like a negative prompt. This approach did not show any clear advantages compared to the negative prompt but at least demonstrated the generality of Predicated Diffusion.

B.3. Automatic Extraction of Propositions

In Experiment (iv), we manually extracted propositions from prompts. However, in most cases, a syntactic dependency parser can be employed to automate this process. To validate this approach, we used spaCy v3.0 to identify the following:

- Nouns for propositions of concurrent existence can be identified as words tagged with NOUN (common noun) and PROPN (proper noun).
- Modifier-noun pairs for propositions of one-to-one correspondence can be identified as word pairs linked by grammatical dependencies, including AMOD (adjectival modifier), NMOD (nominal modifier), COMPOUND (compound nouns) and ACOMP (adjectival complement).
- Possessor-possession pairs for propositions of possession can be identified as subject-object pairs where the verbs indicate possession, namely, 'have,' 'wear,' 'grasp,' and 'hold.'

For example, from the prompt "Woman wearing a black coat holding up a red cellphone," we successfully extracted:

- Nouns: Woman, coat, cellphone
- Modifier-noun pairs: [black, coat], [red, cellphone]
- Possessor-possession pairs: [Woman, coat], [Woman, cellphone]

These results enabled us to automatically create propositions using a simple script.

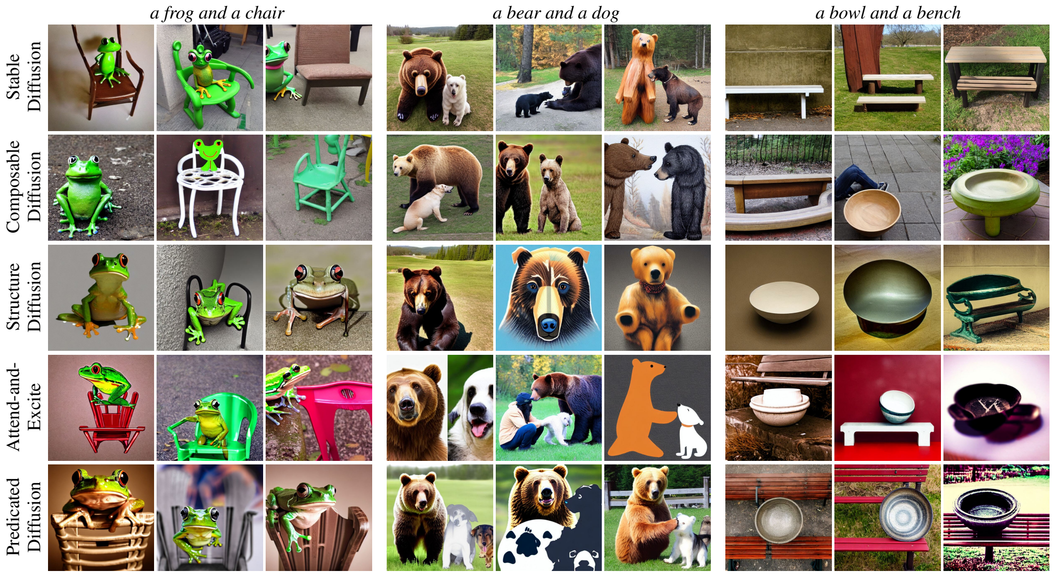


Figure A2. Additional results of Experiment (i) for concurrent existence. See also Fig. 3.

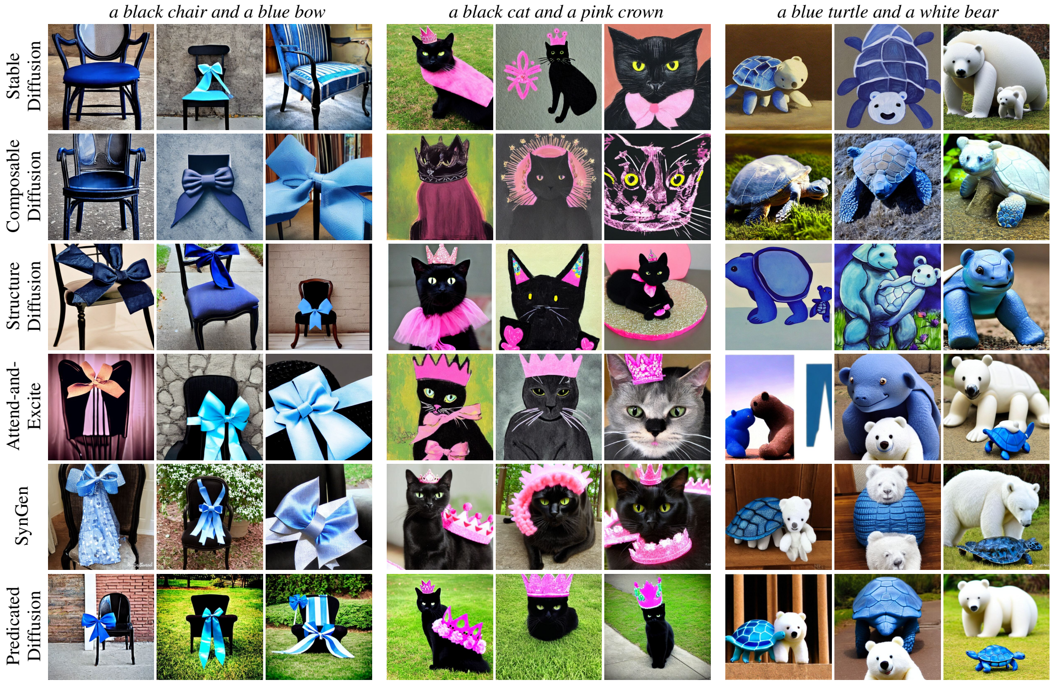


Figure A3. Additional results of Experiment (ii) for one-to-one correspondence. See also Fig. 4.

From six text prompts in Figs. 6 and A5, we successfully extracted all propositions (shown below the images) with two exceptions: the possession relationships in prompts "A black bird with a red beak" and "A white teddy bear with a green shirt and a smiling girl." The preposition "with" expresses possession in these cases but potentially expresses existence in other cases. Given that "with" can have multiple meanings, more advanced syntactic analysis might be necessary. Nonetheless, our findings generally support the sufficiency of simple syntactic analysis.

Furthermore, we believe it is crucial for users to explicitly use predicate logic to clarify their intentions that cannot be fully expressed in text. The meaning of "with," whether indicating a possession relationship or merely concurrent existence, can sometimes be ambiguous even for human readers. By intentionally employing our proposed method,

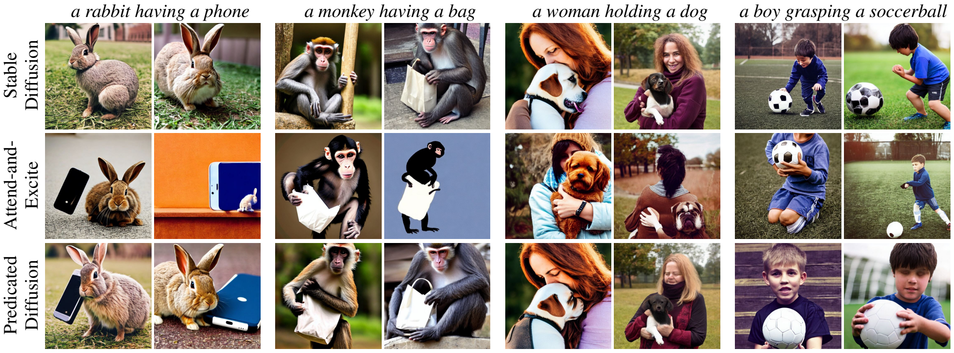


Figure A4. Additional results of Experiment (iii) for possession. See also Fig. 5 in the main body.

Model	Similarity‡	CLIP-IQA
Stable Diffusion	0.353 / 0.699	0.765
SynGen	0.371 / 0.736	0.753
Proposed	**0.387 / 0.745**	**0.777**

‡Text-image similarity and text-text similarity.

Table A4. Results of three objects and three attributes.

	Experiment (i)		Experiment (ii)	
Model	**Similarity‡**	**CLIP-IQA**	**Similarity‡**	**CLIP-IQA**
SynGen	–	–	72.2/70.1	48.9
Proposed	67.4/73.0	55.2	81.2/73.5	54.9

‡Text-image similarity and text-text similarity.

Table A5. Percentages of Improvement.

users can clearly express their intentions and eliminate such ambiguities.

B.4. Additional Analyses

Our proposed Predicated Diffusion incorporates a variety f loss functions, which might raise concerns about achieving suboptimal solutions and thus potentially diminishing the fidelity and quality of generated images. However, the results in Figs. 6, A5, and A6 visually suggest these concerns are unwarranted. For a quantitative evaluation, we generated images of three objects and three attributes, using 15 different loss functions under the same experimental settings as those used in Experiment (ii). The propositions are exemplified in the leftmost column of Fig. A6 The results in Table A4 confirm that our optimization process is robust and successfully handles multiple loss functions without failure. Despite the complexity of the scenarios, objective evaluations through CLIP-IQA rated our method as producing the highest quality images, surpassing those generated by SynGen. Our primary goal is to guide the image generation process rather than strictly fulfill logical constraints, which is why we used fuzzy logic. As such, we view suboptimal results as acceptable within the context of our framework.

Our Predicated Diffusion yielded cartoonish results in the lion/apple case shown in Fig. 3. However, this does not suggest that Predicated Diffusion degrades the naturalness of the generated images. Since the dataset comprises both photos and paintings, what style appears is random. The vanilla Stable Diffusion also generated images of painting style in the cases of bird/cat in Fig. 3, and cat/crown and turtle/bear in Fig. A3; there is no consistent trend across methods. Notably, our proposed method supports the use of style-specifying keywords (like "a photo of X"), which allows users to intentionally choose the image style, as demonstrated in Fig. A9.

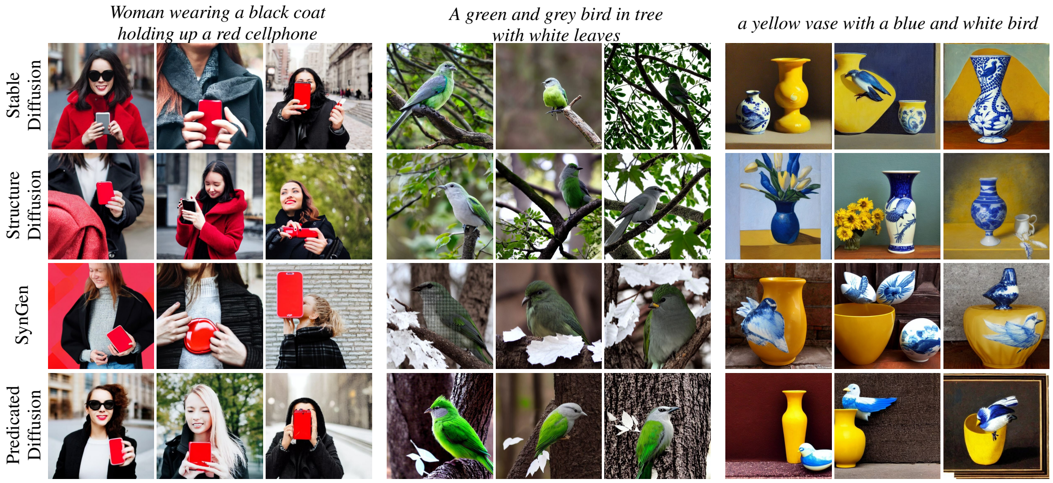
In Table A5, we present the percentage at which each method improves the fidelity and quality of images compared to vanilla Stable Diffusion. Our Predicated Diffusion exhibited statistically significant improvements in all metrics, with $p \ll 0.0001$. Especially, it increased the text-image similarity in 81.2% of 10,000 images generated in Experiment (ii), indicating its consistent efficacy across a broad range of prompts, rather than being limited to particular ones. Moreover, it improved the quality of 55% of images, in contrast to SynGen, which reduced quality in more than 50% of the cases.

B.5. Ablation Study

In Section 3, we used implications to represent the modification by adjectives and the possession of objects as $\forall x.\, Noun(x) \rightarrow Adjective(x)$ and $\forall x.\, Object(x) \rightarrow Subject(x)$, respectively. In this section, we explored the effects of reversing the direction of these implications as an ablation study.

Figure A7 summarizes the results for modification. In the first row, a noun implicates an adjective, indicating that the object specified by the noun is uniformly colored by the hue specified by the adjective.

This implication allows other objects, such as the background, to share the same color, making it suitable for representing modifications in general. As another example, consider the prompt "a sunbathed car," which indicates that the

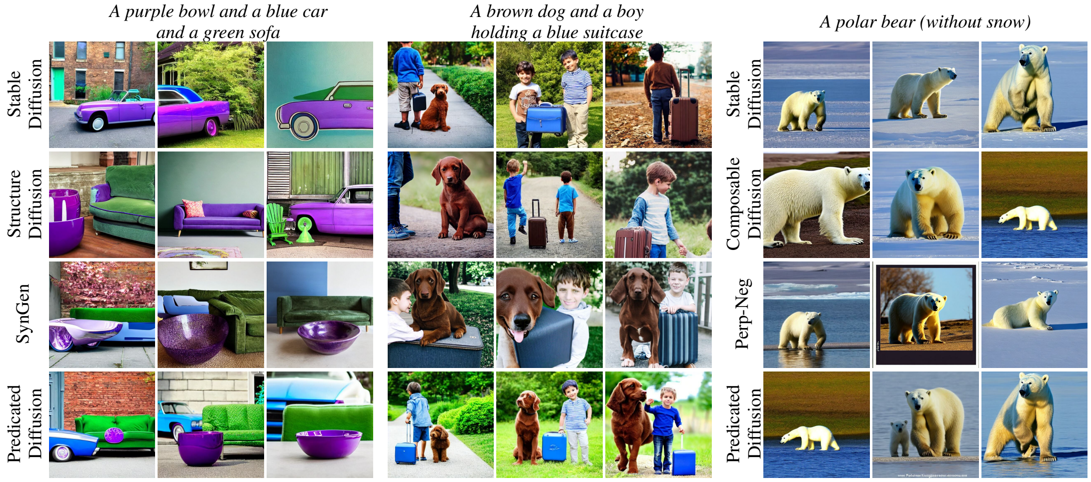

Woman wearing a black coat holding up a red cellphone
A green and grey bird in tree with white leaves
a yellow vase with a blue and white bird
Stable Diffusion
Structure Diffusion
SynGen
Predicated Diffusion
Proposition

$$(\exists x.\, Woman(x)) \land (\exists x.\, Coat(x)) \land (\exists x.\, Cellphone(x)) \land (\forall x.\, Coat(x) \leftrightarrow Black(x)) \land (\forall x.\, Cellphone(x) \leftrightarrow Red(x)) \land (\forall x.\, Coat(x) \rightarrow Woman(x)) \land (\forall x.\, Cellphone(x) \rightarrow Woman(x))$$

$$(\exists x.\, Bird(x)) \land (\exists x.\, Tree(x)) \land (\exists x.\, Leaves(x)) \land (\forall x.\, Bird \leftrightarrow (Grey(x) \lor Green(x))) \land (\forall x.\, Leaves(x) \leftrightarrow White(x))$$

$$(\exists x.\, Vase(x)) \land (\exists x.\, Bird(x)) \land (\forall x.\, Vase(x) \leftrightarrow Yellow(x)) \land (\forall x.\, Bird \leftrightarrow (Blue(x) \lor White(x)))$$

Figure A5. Example results of Experiment (iv) using prompts in ABC-6K. See also Figs. 6 and A6.


A purple bowl and a blue car and a green sofa
A brown dog and a boy holding a blue suitcase
A polar bear (without snow)
Stable Diffusion
Structure Diffusion
SynGen
Predicated Diffusion
Composable Diffusion
Perp-Neg
Proposition

$$(\exists x.\, Bowl(x)) \land (\exists x.\, Car(x)) \land (\exists x.\, Sofa(x)) \land (\forall x.\, Bowl(x) \leftrightarrow Purple(x)) \land (\forall x.\, Car(x) \leftrightarrow Blue(x)) \land (\forall x.\, Sofa(x) \leftrightarrow Green(x)) \land (\forall x.\, Bowl(x) \rightarrow \neg Blue(x)) \land (\forall x.\, Bowl(x) \rightarrow \neg Green(x)) \land (\forall x.\, Car(x) \rightarrow \neg Purple(x)) \land (\forall x.\, Car(x) \rightarrow \neg Green(x)) \land (\forall x.\, Sofa(x) \rightarrow \neg Blue(x)) \land (\forall x.\, Sofa(x) \rightarrow \neg Purple(x))$$

$$(\exists x.\, Dog(x)) \land (\exists x.\, Boy(x)) \land (\exists x.\, Suitcase(x)) \land (\forall x.\, Dog(x) \leftrightarrow Brown(x)) \land (\forall x.\, Suitcase(x) \leftrightarrow Blue(x)) \land (\forall x.\, Suitcase(x) \rightarrow Boy(x))$$

$$(\exists x.\, Bear(x)) \land \neg(\exists x.\, Snow(x))$$

Figure A6. Example results of Experiment (iv) using our original prompts. See also Figs. 6 and A5.

car should be depicted as sunbathed and allows other objects to also appear sunbathed. Conversely, when an adjective implicates a noun, the area colored by the adjective becomes a subset of that of the object, suggesting partial coloring of the object, as shown in the second row. With the biimplication in the last row, the color and object regions perfectly overlap. Therefore, biimplication is preferable to avoid attribute leakage. A clear correspondence can be found between the semantics of propositions and the generated results.

Figure A8 summarizes the results for possession. In the first row, a grammatical object implicates a grammatical

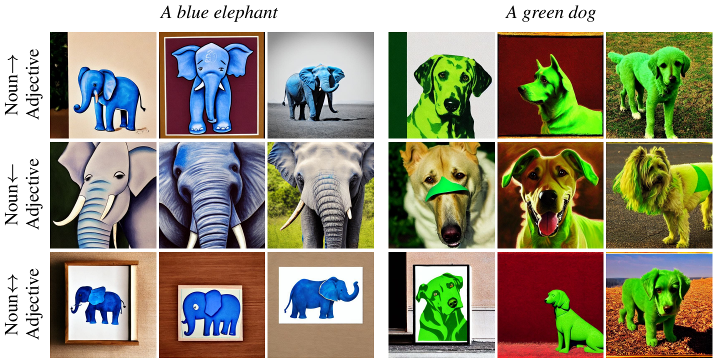


Figure A7. Ablation study for the modification by the implication.

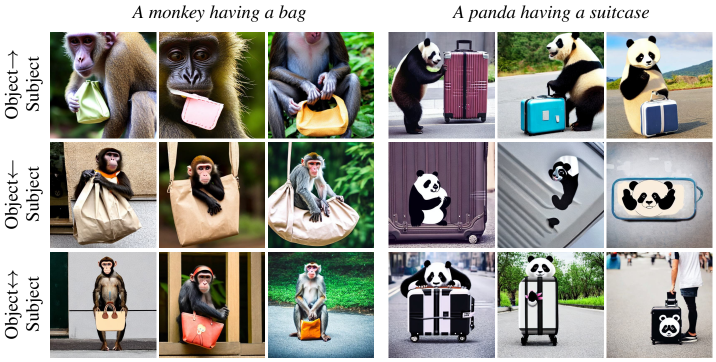
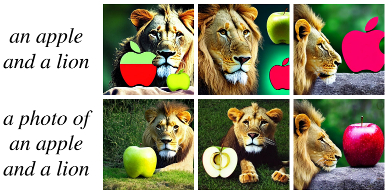


Figure A8. Ablation study for the possession by the implication.

Figure A9. Style-Specifying Keywords.



subject, and the results show the subjects in possession of the objects. Conversely, the second row shows that when the subject implicates the object, the subject becomes part of the object. For example, instead of “A monkey having a bag,” the situation resembles “A bag envelops a monkey.” Similarly, instead of “A panda having a suitcase,” the scene is more like “A panda serves as a pattern on the suitcase.” In cases of biimplication, the subject and object are often mixed together to form one object. Thus, implicating the subject by the object most accurately represents the subject in possession of the object.

B.6. Generality of Challenges

In the main body, we focused on Stable Diffusion v1.4. This is because all comparison methods employed it as their backbone, allowing for fair and accurate comparisons. However, one might raise concerns that the challenges identified are specific to Stable Diffusion alone. To address this and demonstrate the generality of the challenges, we conducted Experiment (iv) on more recently proposed text-to-image diffusion models, Stable Diffusion XL (SDXL) v1.0[3] and Midjourney v5.2[4], that is, we generated images using the prompts as in Figs. 6, A5, and A6. The results are summarized in Fig. A10.

Midjourney and SDXL are clearly superior to Stable Diffusion v1.4 in terms of the quality of the generated images, but they are still prone to challenges discussed in this paper; missing objects, object mixture, attribute leakage, and possession failure. In our first prompt “A black bird with red beak,” Midjourney incorrectly colored the bird’s wings red instead of its beak, exemplifying a case of attribute leakage. SDXL sometimes succeeded but also made the same mistake. For the second prompt “A white teddy bear with a green shirt and a smiling girl,” both SDXL and Midjourney often misidentified the owner of the green shirt as the girl instead of the teddy bear, a typical instance of possession failure. Interestingly, all images generated by Midjourney

---

[3] `https://github.com/Stability-AI/generative-models` (MIT license)

[4] `https://www.midjourney.com/` (Terms of Service: `https://docs.midjourney.com/docs/terms-of-service`)

Table A6. Comparison with Stable Diffusion XL.

	Experiment (i) for Concurrent Existence		Experiment (ii) for One-to-One Correspondence		Experiment (iii) for Possession	
Methods	**Similarity‡**	**CLIP-IQA**	**Similarity‡**	**CLIP-IQA**	**Similarity‡**	**CLIP-IQA**
Stable Diffusion	0.326 / 0.767	0.761	0.345 / 0.744	0.756	0.320 / 0.811	0.762
Stable Diffusion + Predicated Diffusion	**0.348 / 0.825**	0.775	**0.379 / 0.811**	0.769	0.345 / 0.855	0.765
Stable Diffusion XL	0.340 / 0.820	**0.777**	0.369 / 0.793	**0.773**	**0.353 / 0.862**	**0.780**

†Using the lenient and strict criterions. ‡Text-image similarity and text-text similarity.

are remarkably similar in layout, suggesting the limitation of diversity. In the third prompt “A baby with green hair laying in a black blanket next to a teddy bear,” Midjourney altered the color of the teddy bear or blanket to green, rather than the baby’s hair. SDXL often successfully gave the baby green hair, but still dyed the blanket or teddy bear green as well. With the fourth prompt “Woman wearing a black coat holding up a red cellphone,” both SDXL and Midjourney often incorrectly switched the colors of the coat and cellphone. The fifth prompt “A green and grey bird in a tree with white leaves,” resulted in both SDXL and Midjourney rarely depicting white leaves. Midjourney often substituted them with white flowers, and the bird was not distinctly grey. For the sixth prompt “A yellow vase with a blue and white bird,” both SDXL and Midjourney consistently mixed object colors. In the seventh prompt “A purple bowl and a blue car and a green sofa,” the bowl was often missing. In Midjourney, the car was often missing as well, with only something resembling headlights attached to the sofa, suggesting the object mixture. The colors were also incorrect in most cases. Finally, in the eighth prompt “A brown dog and a boy holding a blue suitcase,” the boy was typically depicted placing the suitcase on the ground, another case of possession failure. In Midjourney, the suitcase was rarely blue, often replaced by the boy’s clothes being blue. Additionally, the boy and dog appeared almost identical in each trial, highlighting a lack of diversity.

In addition, we conducted Experiments (i)–(iii) on SDXL as well, evaluating similarities and CLIP-IQA, with the results summarized in Table A6. Note that, even when using the same random seeds, the images generated by Stable Diffusion and SDXL are totally different due to the differences in image resolution and network structure. In terms of image quality (CLIP-IQA), SDXL consistently outperforms both Stable Diffusion and Predicated Diffusion, as also observed in Fig. A10. However, in terms of fidelity (similarity), Predicated Diffusion with Stable Diffusion as its backbone surpasses SDXL in Experiments (i) and (ii). SDXL has approximately three times the number of parameters as Stable Diffusion and incorporates various improvements, such as an additional text encoder and pooled text embeddings. Moreover, it has been trained on a more extensive dataset. Despite these enhancements, Predicated Diffusion proves to be more effective in preventing missing objects and attribute leakage. In Experiment (iii), while SDXL outperforms Predicated Diffusion, the margin is smaller than the improvement achieved by Predicated Diffusion over vanilla Stable Diffusion.

Our findings demonstrate that Predicated Diffusion effectively overcomes a variety of challenges. While cutting-edge models excel in generating high-quality images, they still struggle with these challenges. Furthermore, the fundamental concept of Predicated Diffusion has the potential to improve these models, providing a promising direction for future research.

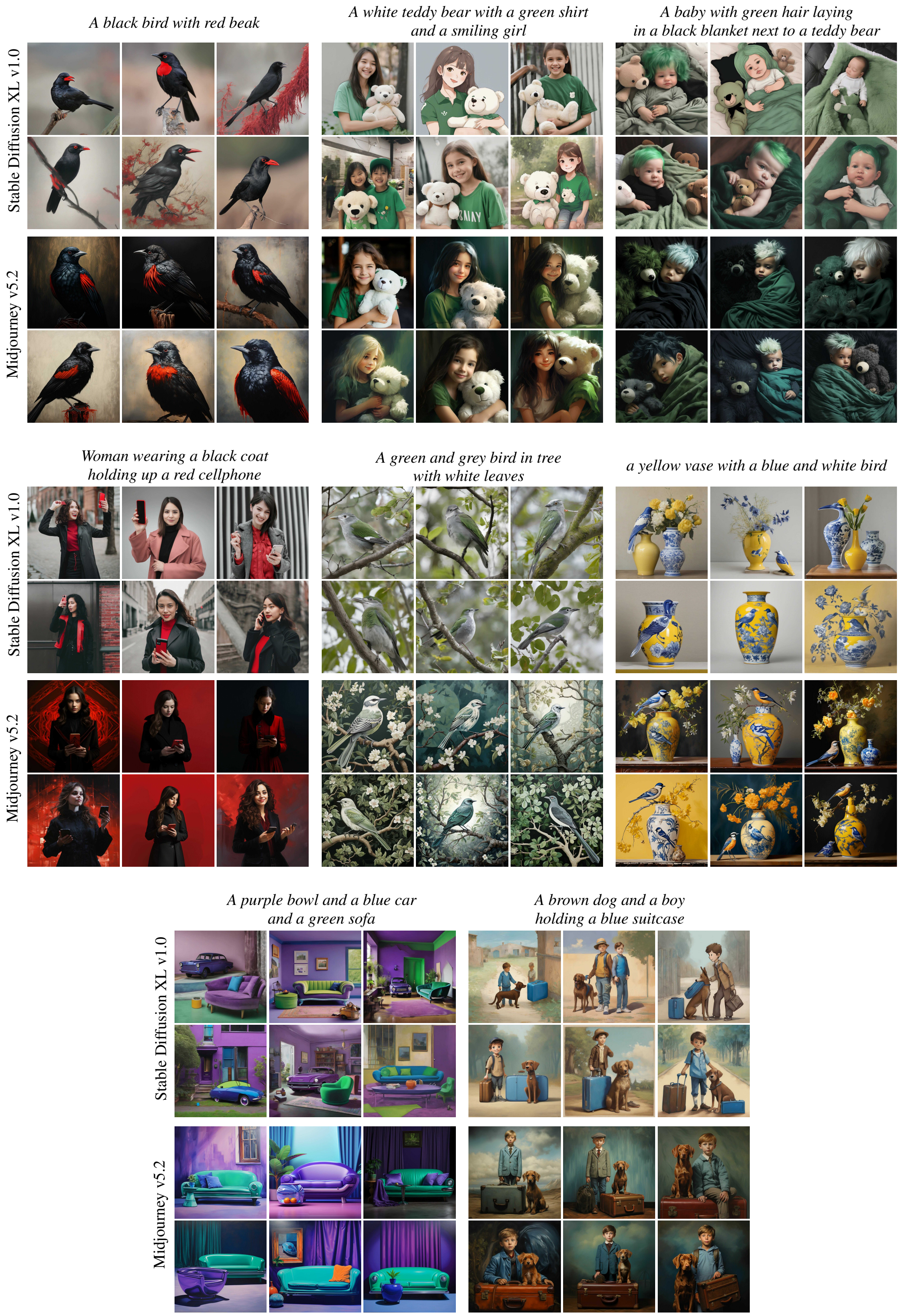

A black bird with red beak
A white teddy bear with a green shirt and a smiling girl
A baby with green hair laying in a black blanket next to a teddy bear
Stable Diffusion XL v1.0
Midjourney v5.2
Woman wearing a black coat holding up a red cellphone
A green and grey bird in tree with white leaves
a yellow vase with a blue and white bird
Stable Diffusion XL v1.0
Midjourney v5.2
A purple bowl and a blue car and a green sofa
A brown dog and a boy holding a blue suitcase
Stable Diffusion XL v1.0
Midjourney v5.2

Figure A10. Examination of more recently proposed text-to-image diffusion models.