LoRAverse: A Submodular Framework to Retrieve Diverse Adapters for Diffusion Models

Supplementary Material

7. Limitations

Although LoRAverse is able to select meaningful adapters for each extracted concept, possible errors during clustering of adapters prior to applying submodular selection may lead to over-retrieval of similar adapters or neglect of relevant ones. If alike adapters were placed in the different clusters or dissimilar adapters are clustered together, rendered diversity could be limited. Moreover, retrieved LoRA models, which could inadvertently amplify societal biases present in training data. Therefore, we advocate for the implementation of safeguards to mitigate both bias and misuse risks.

8. Proof of Submodularity

To justify the use of a greedy algorithm for optimizing Eq. 4, we formally prove that the combined objective function $\mathcal{F}(\mathcal{P})$ is submodular. We begin by analyzing the components $\mathcal{F}_{\text{rel}}(\mathcal{P})$[2] and $\mathcal{F}_{\text{div}}(\mathcal{P})$[3] separately.

Lemma 1 *$\mathcal{F}_{rel}(\mathcal{P})$, formulated in Eq. 2, is a modular function and therefore submodular.*
Proof. A function is modular if it can be expressed as a linear sum over its elements, i.e., $\mathcal{F}(\mathcal{P}) = \sum_{v\in\mathcal{P}} w(v)$, for some function $w : \mathcal{V} \rightarrow \mathbb{R}$. Here, $\mathcal{F}_{rel}(\mathcal{P})$ is modular with $w(a_i) = \mathcal{F}_{sim}(\phi(a_i), \phi(s))$. For any $\mathcal{P} \subseteq \mathcal{V}$ and $v \notin \mathcal{P}$, the marginal gain of adding v is:

$$\mathcal{F}_{rel}(\mathcal{P} \cup \{v\}) - \mathcal{F}_{rel}(\mathcal{P}) = \mathcal{F}_{sim}(\phi(a_i), \phi(s)) \tag{7}$$

This marginal gain is independent of $\mathcal{P}$, satisfying the submodularity condition with equality:

$$\mathcal{F}_{rel}(\mathcal{P} \cup \{v\}) - \mathcal{F}_{rel}(\mathcal{P}) = \mathcal{F}_{rel}(\mathcal{R} \cup \{v\}) - \mathcal{F}_{rel}(\mathcal{R}) \tag{8}$$

for all $R \subseteq P$. Thus, $\mathcal{F}_{rel}(\mathcal{P})$ is modular and therefore submodular. □

Lemma 2 *$\mathcal{F}_{div}(\mathcal{P})$, formulated in Eq. 3, is submodular.*
Proof. Suppose $v \in \mathcal{C}_k$ for some cluster k. Since the clusters $\{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ are disjoint, adding v only affects the term for $\mathcal{C}_k$ in the sum. For all other clusters $j \neq k$, the marginal gain is zero. Thus, the inequality reduces to:

$$\begin{aligned}\log(1+\mathcal{S}_\mathcal{P}+\mathcal{F}_{rew}(v)) - \log(1+\mathcal{S}_\mathcal{P}) \leq \\ \log(1+\mathcal{S}_\mathcal{R}+\mathcal{F}_{rew}(v)) - \log(1+\mathcal{S}_\mathcal{R})\end{aligned} \tag{9}$$

where $\mathcal{S}_\mathcal{P} = \sum_{a_i\in\mathcal{C}_k\cap\mathcal{P}} \mathcal{F}_{rew}(a_i)$[4] and $\mathcal{S}_\mathcal{R} = \sum_{a_i\in\mathcal{C}_k\cap\mathcal{R}} \mathcal{F}_{rew}(a_i)$. The function $g(x) = \log(1+x)$ is concave, so its derivative $g'(x) = 1/(1+x)$ is decreasing. By the Mean Value Theorem, for some $\xi \in [S_\mathcal{R}, S_\mathcal{R} + \mathcal{F}_{rew}(v)]$ and $\zeta \in [S_\mathcal{R}, S_\mathcal{R} + \mathcal{F}_{rew}(v)]$:

$$\begin{aligned}\log(1+\mathcal{S}_\mathcal{R}+\mathcal{F}_{rew}(v)) - \log(1+\mathcal{S}_\mathcal{R}) = \frac{\mathcal{F}_{rew}(v)}{1+\xi} \\ \log(1+\mathcal{S}_\mathcal{P}+\mathcal{F}_{rew}(v)) - \log(1+\mathcal{S}_\mathcal{P}) = \frac{\mathcal{F}_{rew}(v)}{1+\zeta}\end{aligned} \tag{10}$$

Since $\mathcal{R} \subseteq \mathcal{P}$, we have $S_\mathcal{R} \leq S_\mathcal{P}$, which implies $\xi \leq \zeta$. Therefore:

$$\frac{\mathcal{F}_{rew}(v)}{1+\zeta} \leq \frac{\mathcal{F}_{rew}(v)}{1+\xi} \tag{11}$$

This establishes the inequality. Since the inequality holds for every cluster $\mathcal{C}_k$, summing over all k preserves submodularity. Thus, $\mathcal{F}_{div}(\mathcal{P})$ is submodular. □

Lemma 3 *$\mathcal{F}_(\mathcal{P})$, formulated in Eq. 4, is submodular.*
Proof. Since $\mathcal{F}(\mathcal{P})$ is a non-negative linear combination of submodular functions, it is itself submodular. □

9. Additional Qualitative Results

Additional qualitative results are in Fig. 13 and Fig. 14, demonstrating LoRAverse's effectiveness in generating diverse and relevant image sets. Our approach balances diversity with image-text alignment, thanks to the clustering-based retrieval process. By selecting adapters from the clusters, our method ensures the generated images reflect both the prompt's content and stylistic variety, offering a richer interpretation compared to other methods.

10. Additional Quantitative Results

To contextualize LoRAverse against stronger yet straightforward alternatives, we add three baselines: RTop-100 and RTop-500, which rank all LoRA models by global prompt similarity and randomly sample $K = 8$ from the top 100 or 500, and Random, which samples from the full pool. As

[2] $\mathcal{F}_{\text{rel}}$ is the abbreviated form of $\mathcal{F}_{\text{relevance}}$.

[3] $\mathcal{F}_{\text{div}}$ is the abbreviated form of $\mathcal{F}_{\text{diversity}}$.

[4] $\mathcal{F}_{\text{rew}}$ is the abbreviated form of $\mathcal{F}_{\text{reward}}$

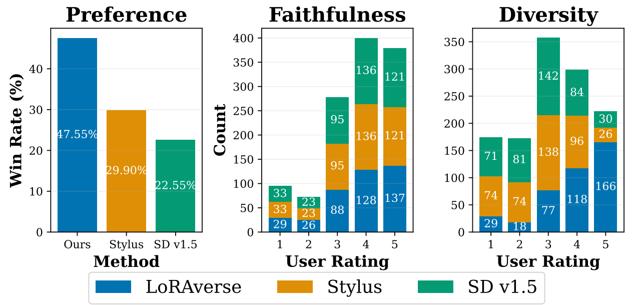


Figure 10. **User Study.** Users were asked between LoRAverse, Stylus, and SD v1.5 which produced preferred outputs, and also to rate the faithfulness and diversity of outputs on a scale of 1 to 5. LoRAverse exhibits both strong desirability and outperforms in diversity.

	CLIP (↑)	TCE (↑)	TIE (↑)	I2I (↓)
LoRAverse	24.94 (14.6%)	**25.72** (2.3%)	**43.04** (0.5%)	**0.719** (-5.8%)
SD v1.5	**25.29** (16.2%)	23.65 (-6.0%)	41.86 (-2.3%)	0.774 (1.4%)
RTop-100	24.99 (14.8%)	25.16 (0.0%)	42.40 (-1.3%)	0.740 (-3.0%)
RTop-500	24.48 (12.5%)	25.05 (-0.4%)	42.35 (-1.1%)	0.739 (-3.2%)
Random	21.76	25.15	42.84	0.763
5-Cluster	24.55	26.10	43.22	**0.716**
10-Cluster	**24.94**	25.72	43.04	0.719
25-Cluster	24.45	**26.39**	**43.41**	0.707
1-Concept	24.73	23.27	40.25	0.772
2-Concept	24.87	**24.02**	40.49	0.763
3-Concept	**24.92**	23.54	**41.61**	**0.762**

Table 3. **Additional Quantitative Comparison.** Results from the additional metrics and the ablation study on the number of clusters and concepts.

summarized in Table 3, LoRAverse matches RTop-100 in CLIP alignment (24.94 vs. 24.99) while achieving higher diversity on all metrics and lower image-to-image similarity. The advantage widens for RTop-500 and Random, whose looser sampling selects more off-topic adapters, leading to lower CLIP scores. These results confirm that the proposed submodular retrieval offers a superior balance between diversity and text-adherence.

11. Ablation Study Over the Number of Clusters and Concepts

We evaluated LoRAverse's sensitivity to the number of clusters formed per concept (5, 10, 25) and the number of concepts extracted from the prompt (1–3). Varying the size of the cluster effectively leaves all metrics stable, that is, it changes all metrics by less than 1.5 points and changes the pairwise similarity by less than 0.02, showing no consistent trend (see Table 3). This indicates LoRAverse is robust to these settings.

12. Details of Concept Extractor

A complete example prompt for the *concept extractor* is provided in the first column of Table 4, utilizing the `gpt-4o-mini` model. We design a structured prompt to help the LLM effectively identify distinct, orthogonal concepts while merging those that are semantically similar. This module is implemented using LangChain. [5].

13. Details of Adapter Safety Checker

A complete input for the *adapter safety checker* is provided in the second column of Table 4, utilizing the `gpt-4o` model. This module ensures ethical integrity by filtering adapters associated with inappropriate, sexual, or anthropomorphic material. In large-scale text-to-image pipelines with diverse LoRA models, there exists a risk of retrieving adapters that could produce harmful outputs. The safety checker mitigates this by evaluating adapter descriptions and removing those violating content guidelines.

The checker is implemented through a structured prompt that helps the LLM identify inappropriate adapters unless explicitly requested by the user. It filters out two main types: (1) potentially sexual content, including nudity or explicit elements, and (2) anthropomorphic content, such as animal-inspired humanoids. This approach ensures only contextually appropriate adapters are selected.

The decision to filter anthropomorphic content stems from observations that these styles often blur fantasy and reality in problematic ways. These figures frequently appear in exaggerated forms that may contribute to objectification of human-like features. Excluding these adapters by default helps maintain more neutral outputs unless specifically requested. Implemented using LangChain [5], the safety checker introduces computational overhead but remains a practical component. This approach balances creative diversity with safety guards, reducing the risk of generating unsafe outputs while preserving quality and diversity in the results.

14. Details of VLM-as-a-Judge

The complete prompt used for evaluating the diversity, quality, and textual alignment of the image sets with `gpt-4o` is outlined in Table 5. Our methodology involves presenting three distinct image sets and using multi-turn prompting techniques to differentiate between them. Each set is numbered sequentially, and the VLM is asked to evaluate their diversity, quality, and textual alignment. The prompt includes a structured rubric, clear instructions, reminders, and example model outputs. To quantify the metrics, the model uses Chain-of-Thought reasoning to assign scores ranging from 0 to 2, based on methodologies from Stylus [9, 40, 42].

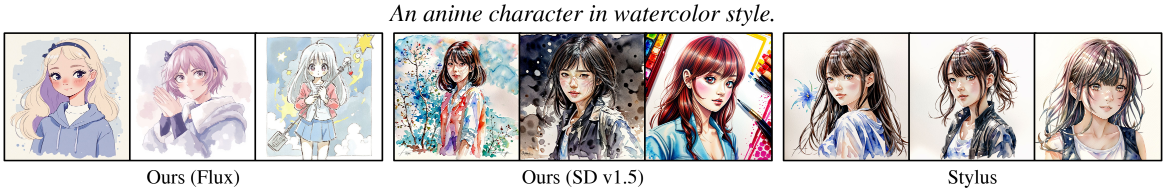
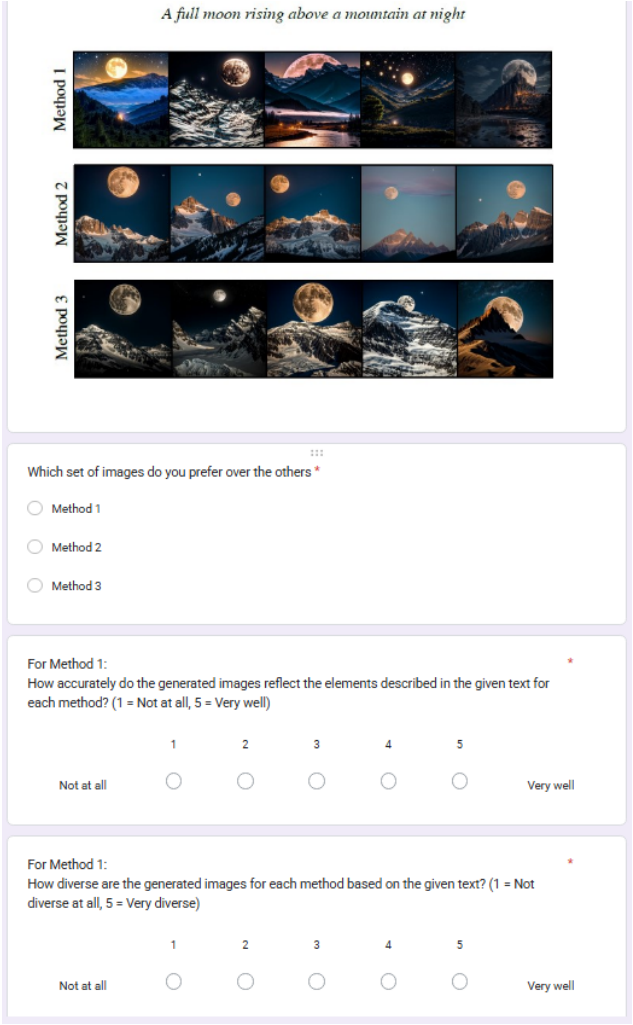


Figure 11. **Compatibility with Flux.** Additional qualitative outputs generated by Flux, LoRAverse, and Stylus.



A full moon rising above a mountain at night

Method 1

Method 2

Method 3

Which set of images do you prefer over the others *

○ Method 1

○ Method 2

○ Method 3

For Method 1: *
How accurately do the generated images reflect the elements described in the given text for each method? (1 = Not at all, 5 = Very well)

	1	2	3	4	5	
Not at all	○	○	○	○	○	Very well

For Method 1: *
How diverse are the generated images for each method based on the given text? (1 = Not diverse at all, 5 = Very diverse)

	1	2	3	4	5	
Not at all	○	○	○	○	○	Very well

Figure 12. **Screenshot of the User Study.**

15. Details of Clustering Implementation

To enhance reproducibility, we provide a detailed overview of our clustering methodology. We use the BERTopic [13] framework with UMAP [28] and HDBSCAN [27] for adapter grouping. The UMAP configuration is set to 5 neighbors, 5 components, and cosine distance metric, prioritizing local structure preservation while reducing dimensionality efficiently. For clustering, HDBSCAN is configured with a minimum cluster size of 3, allowing for the formation of micro-clusters. We use euclidean distance in the low-dimensional space and enable prediction data to support unseen data inference. The number of clusters is not predetermined but emerges dynamically from the density-based clustering, ensuring adaptability across datasets. Additionally, we set the number of topics to 10. These design choices capture the underlying structure of LoRA models.

16. Details of User Study

For each prompt, we presented a set of 5 images for LoRAverse, Stylus, and SD v1.5 and asked 3 questions:

- **Question 1 - Preference**: ”Which set of images do you prefer over the others”
- **Question 2 - Faithfulness**: ”For [GIVEN METHOD]: How accurately do the generated images reflect the elements described in the given text for each method? (1 = Not at all, 5 = Very well)”
- **Question 3 - Diversity**: ”For [GIVEN METHOD]: How diverse are the generated images for each method based on the given text? (1 = Not diverse at all, 5 = Very diverse)”

Fig. 10 shows the win-rate (proportion of times a user chose a method as their preferred), along with the distribution of user ratings for each method for faithfulness and diversity. We observe that while users gave relatively similar ratings for the image faithfulness, there are significantly more users rating LoRAverse with higher diversity. A screenshot of the survey for a sample prompt and question can be seen in Fig. 12.

17. Compatibility with Other Text-to-Image Models

LoRAverse is inherently backbone-agnostic because it represents each adapter solely by its CLIP embedding, the submodular retrieval objective–and thus the entire pipeline–remains unchanged when the underlying text-to-image model is swapped. To validate this claim, we indexed 300 publicly released Flux-compatible LoRA adapters from Hugging Face [1], created their embeddings, and ran our selection algorithm using Flux.1 backbone (see Fig. 11).

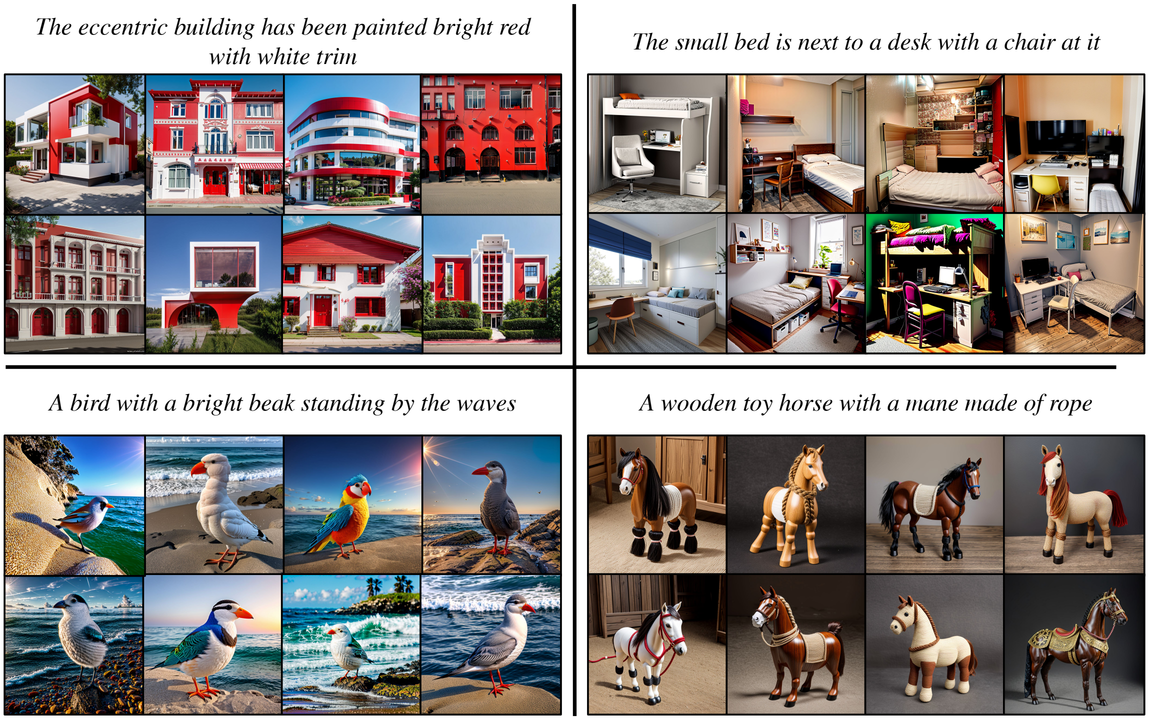

The eccentric building has been painted bright red with white trim
The small bed is next to a desk with a chair at it
A bird with a bright beak standing by the waves
A wooden toy horse with a mane made of rope

Figure 13. **Additional Qualitative Results.** Additional LoRAverse outputs showcasing diverse image sets generated by retrieving various LoRA models while maintaining relevance to the user prompts.

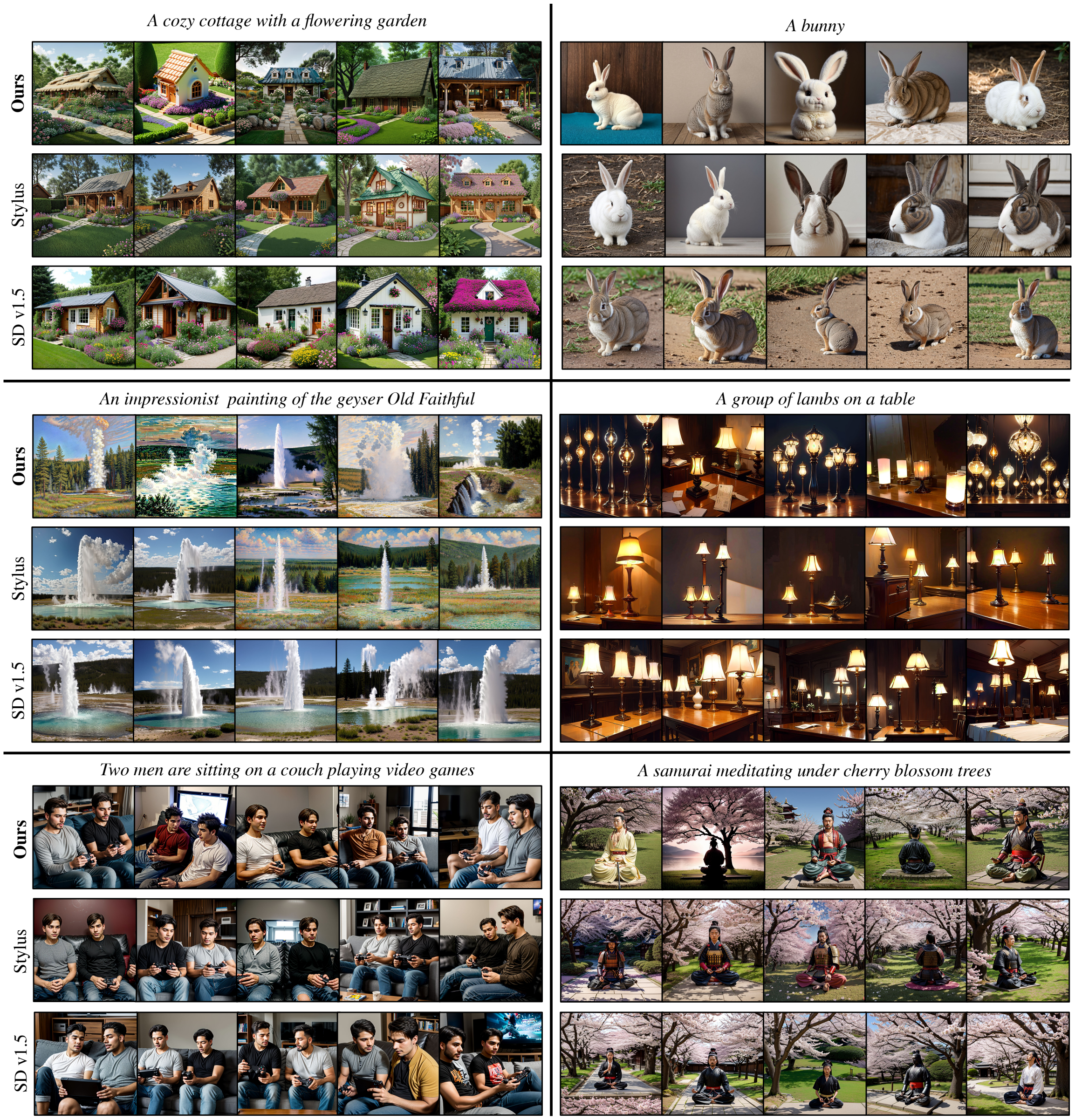

A cozy cottage with a flowering garden
A bunny
An impressionist painting of the geyser Old Faithful
A group of lambs on a table
Two men are sitting on a couch playing video games
A samurai meditating under cherry blossom trees
Ours
Stylus
SD v1.5

Figure 14. **Additional Qualitative Comparison.** Additional outputs generated by LoRAverse, Stylus, and SD v1.5 demonstrating diverse image sets.

System Prompt You are a specialist to extract meaningful keywords and provide explanations for prompts designed for text-to-image models. Your task is to identify keywords or phrases that significantly impact the visual content generated by such models.	**System Prompt** You are a specialized adapter content filter for text-to-image pipelines. Your primary task is to identify and filter out adapters that contain inappropriate content based on the following criteria: 1) Potentially Sexual Content: Descriptions referencing nudity, sexualized poses, or explicit elements. 2) Anthropomorphic Content: Descriptions featuring catgirls, animal-inspired humanoids, or similar anthropomorphic characters. Only exclude adapters containing these types of content unless the prompt explicitly mentions or requires them. You must carefully evaluate the descriptions of the adapters and flag those that violate the criteria. For each filtered adapter, provide its index and a concise explanation for why it was flagged. Your output should strictly follow the specified JSON format provided in the main prompt, listing the flagged adapters along with their explanations.
User Prompt You are a keyword extractor specialized in identifying concepts and their descriptive adjectives from prompts used in text-to-image models. Your task is to detect key concepts (keywords with their adjectives) that significantly influence the content, style, or composition of the generated image. Ensure that the extracted concepts are relevant to the visual elements of the scene and describe distinct entities or attributes. Do not split related concepts that describe a single entity. Avoid extracting verbs describing actions. Key Instructions: 1) Extract descriptive concepts: Always extract keywords that include any relevant descriptive adjectives or qualifiers that modify them in the prompt. If adjectives or modifiers are provided, ensure they are part of the concept, as they provide crucial detail to the entity. For example: - For the prompt ”a large, ancient stone castle,” return the keyword ”large, ancient stone castle” rather than just ”castle.” The adjectives ”large,” ”ancient,” and ”stone” provide important context and description that shape the entity’s appearance and character. - For the prompt ”a vibrant painting of a tropical sunset,” return the keyword ”vibrant painting of a tropical sunset,” as ”vibrant” enhances the concept of the painting and ”tropical” adds a layer of specificity to the sunset. 2) Focus on impactful concepts: Keywords should describe specific entities, concepts, styles, or attributes central to the generated image. Avoid overly general terms (e.g., ”thing” or ”place”). Ignore verbs describing an action.” 3) Combine related concepts: If multiple keywords collectively describe a single entity, treat them as one unified concept. Do not split descriptive phrases or components that together define the same object, scene, or idea. If extracted keywords simply describe or qualify a specific entity, they must be merged into a single concept. For example: - In the prompt ”a photo of San Francisco’s Golden Gate Bridge,” the entire phrase should be treated as one concept, ”a photo of San Francisco’s Golden Gate Bridge,” because all parts describe a single entity—the photo. - In the prompt ”a picture of some food on the plate,” the entire phrase should be treated as one concept, ”a picture of some food on the plate,” because all parts collectively describe the picture and its content. 4) Provide explanations: For each keyword, provide a concise explanation of why it was chosen, focusing on its role in shaping the image’s appearance, theme, or composition. 5) Stick to the specified format: The output must follow the JSON format provided below within <output_format> and </output_format> tags. <output_format> [{ ”explanation”: ”[Reason for selecting this keyword]”, ”keyword”: ”[Keyword identified]” }, { ”explanation”: ”[Reason for selecting this keyword]”, ”keyword”: ”[Keyword identified]” }, ...] </output_format> Extract and explain keywords for the following prompt enclosed within <prompt> and </prompt> tags: <prompt> {prompt} </prompt> Provide your response in the JSON format specified above without the <output_format> and “‘json tags.	**User Prompt** You are an adapter content filter for text-to-image models. Adapters are plug-in modules that enhance image quality in specific domains, as described in their descriptions. Your task is to identify adapters that are inappropriate for use in the context of the provided prompt and keyword. Specifically, filter out adapters whose descriptions contain the following types of content, unless the prompt explicitly mentions them: 1) Potentially Sexual Content: Descriptions referencing nudity, sexualized poses, or other explicit elements. 2) Anthropomorphic Content: Descriptions featuring catgirls, animal-inspired humanoids, or similar anthropomorphic characters. You are provided with adapter descriptions, each marked with an index value. Your response should list the indices of the inappropriate adapters, along with a brief explanation for why they were filtered. Key Instructions: 1) Filter Content Types: Remove adapters containing the specified types of content unless the prompt directly mentions or explicitly requires them. 2) Provide Explanations: For each filtered adapter, include a concise explanation of why it was flagged, referencing the content type(s) it violates. 3) Follow the output format: Use the JSON format provided below, enclosed within <output_format> and </output_format> tags. <output_format> [{ ”explanation”: ”[Reason for why this adapter is inappropriate]” ”id”: ”[id of the inappropriate adapter]” }, { ”explanation”: ”[Reason for why this adapter is inappropriate]” ”id”: ”[id of the inappropriate adapter]” }, ...] </output_format> Task Section: Filter the inappropriate adapters for the provided keyword and prompt enclosed below: <keyword_and_prompt> Keyword: {keyword} Prompt: {prompt} </keyword_and_prompt> The adapter descriptions to be filtered are enclosed below: <adapter_descriptions> {adapter_descriptions_str} </adapter_descriptions> Provide your response in the JSON format specified above without the <output_format> and “‘json tags.

Table 4. Complete prompts for the *concept extractor* and *adapter safety checker* respectively.

System Prompt You are a precise and objective Photoshop expert tasked with evaluating the diversity of three given image sets. Your role is to analyze and score the diversity of these sets based on predefined criteria. You must provide a clear decision on which image set is more diverse, along with detailed explanations for your reasoning. Your assessment should be factual, concise, and unbiased, following the specified JSON format. Scoring Criteria: Diversity scores are assigned as follows: - 2 (Very diverse): The image set displays significant variation across themes and main subjects. - 1 (Somewhat diverse): The set shows some diversity but lacks variation in either theme interpretation or main subjects. - 0 (Not diverse): The set contains minimal or no variation in both theme interpretation and main subjects. Diversity Evaluation Breakdown: You will assess diversity based on two key aspects: 1) Theme Interpretation: The theme should exhibit multiple interpretations. For example, if the theme is ”It’s raining cats and dogs”, a diverse set should include both literal (cats and dogs falling from the sky) and figurative (heavy rain) representations. If the set only includes images of heavy rain or only of animals, it should receive a score of 1 instead of 2. 2) Main Subject: The diversity score should reflect changes in the primary subject of the images. For example, a set containing images of apples and children dressed as apples is more diverse than a set with only children dressed as apples. A set with varying focal points across different images should receive a higher diversity score.	**System Prompt** You are a precise and objective Photoshop expert tasked with evaluating the composition quality of three given image sets. Your role is to analyze and score the quality of these sets based on predefined criteria. You must provide a clear decision on which image set has higher quality, along with detailed explanations for your reasoning. Your assessment should be factual, concise, and unbiased, following the specified JSON format. Scoring Criteria: Compositional quality scores are assigned as follows: - 2 (Very quality): The image set displays high quality. - 1 (Somewhat quality): The image set is visually aesthetic but has elements with distortion, missing, or extra features. - 0 (Not quality): The set contains minimal or no quality. Quality Evaluation Breakdown: You will assess quality based on three key aspects: 1) Clarity: Score 0 if the image is blurry, poorly lit, or has poor composition with objects obstructing each other. 2) Disfigured Parts: This applies to both body parts of humans/animals and objects like motorcycles. Score 0 if parts are severely disfigured such as fingers showing lips and teeth warped in. Score 1 for minor anatomical errors like a hand with 6 fingers. 3) Detail: Score 0 for the appearance inconsistent with the environment. Score 1 for acceptable but basic detail such as monochrome and flat surfaces. Score 2 for rich, realistic detail like a sailboat showing dynamic ripples and ornate patterns.	**System Prompt** You are a precise and objective Photoshop expert tasked with evaluating the textual alignment of three given image sets based on the provided prompt. Your role is to analyze and score the textual alignment of these sets according to the following criteria. You must provide a clear decision on which image set aligns best with the prompt, along with detailed explanations for your reasoning. Your assessment should be factual, concise, and unbiased, following the specified JSON format. Scoring Criteria: Textual alignment scores are assigned as follows: - 2 (Fully aligned): The image set displays high textual alignment with the prompt. - 1 (Somewhat aligned): The set incorporates part of the theme but not all elements are correctly represented. - 0 (Not aligned): The set contains minimal or no textual alignment with the prompt. Here are some examples: - If the prompt is ”shoes” and an image depicts a sock, the score would be 0 (not aligned). - If the prompt is ”shoes without laces” but the image shows shoes with laces, the score would be 1 (somewhat aligned). - If the prompt is ”a concert without fans,” but an image includes fans, select the set with fewer fans. This would be scored based on the image set that most closely matches the prompt.
User Prompt This is one of the image sets. Please reply ’ACK’. <image_set_1>	**User Prompt** This is one of the image sets. Please reply ’ACK’. <image_set_1>	**User Prompt** This is one of the image sets. Please reply ’ACK’. <image_set_1>
Assistant ACK	**Assistant** ACK	**Assistant** ACK
User Prompt This is one of the image sets. Please reply ’ACK’. <image_set_2>	**User Prompt** This is one of the image sets. Please reply ’ACK’. <image_set_2>	**User Prompt** This is one of the image sets. Please reply ’ACK’. <image_set_2>
Assistant ACK	**Assistant** ACK	**Assistant** ACK
User Prompt This is one of the image sets. Please reply ’ACK’. <image_set_3>	**User Prompt** This is one of the image sets. Please reply ’ACK’. <image_set_3>	**User Prompt** This is one of the image sets. Please reply ’ACK’. <image_set_3>
Assistant ACK	**Assistant** ACK	**Assistant** ACK

User Prompt
Rate the diversity of the three provided image sets using the scoring criteria above. For each group, assign each set a diversity score along with a detailed explanation in the following JSON output format:
 	
JSON Output Format: [
{
”image_set_1_explanation”: #Your detailed evaluation of the diversity in Image Set 1#,
”image_set_1_score”: #2, 1, or 0#
},
{
”image_set_2_explanation”: #Your detailed evaluation of the diversity in Image Set 2.#,
”image_set_2_score”: #2, 1, or 0#
},
{
”image_set_3_explanation”: #Your detailed evaluation of the diversity in Image Set 3.#,
”image_set_3_score”: #2, 1, or 0#
},
{
”preference_explanation”: #Your reasoning for choosing the more diverse set.#,
”choice”: #IMAGE_SET_1, IMAGE_SET_2, or IMAGE_SET_3#
}
]
 	
I will make my own judgement using your results, your response is just an opinion as part of a rigorous process. I provide additional requirements below:
- Do not forget to reward different main subjects in the diversity score.
- You must pick a group for ”More Diversity,” neither is not an option.
- If the decision is close, make a choice and clarify your reasoning.
 	
Provide your response directly in the specified JSON format without “‘json tags.
	User Prompt
	Rate the quality of the three provided image sets using the scoring criteria above. For each group, assign each set a quality score along with a detailed explanation in the following JSON output format:
	
	JSON Output Format: [
	{
	”image_set_1_explanation”: #Your detailed evaluation of the quality in Image Set 1#,
	”image_set_1_score”: #2, 1, or 0#
	},
	{
	”image_set_2_explanation”: #Your detailed evaluation of the quality in Image Set 2.#,
	”image_set_2_score”: #2, 1, or 0#
	},
	{
	”image_set_3_explanation”: #Your detailed evaluation of the quality in Image Set 3.#,
	”image_set_3_score”: #2, 1, or 0#
	},
	{
	”preference_explanation”: #Your reasoning for choosing the higher quality set.#,
	”choice”: #IMAGE_SET_1, IMAGE_SET_2, or IMAGE_SET_3#
	}
	]
	
	I will make my own judgement using your results, your response is just an opinion as part of a rigorous process. I provide additional requirements below:
	- You must pick a group for ”Better Quality,” neither is not an option.
	- If the decision is close, make a choice and clarify your reasoning.
	
	Provide your response directly in the specified JSON format without “‘json tags.
		User Prompt
		Rate the textual alignment of the three provided image sets using the scoring criteria above. For each group, assign each set a textual alignment score along with a detailed explanation in the following JSON output format:
		JSON Output Format: [
		{
		”image_set_1_explanation”: #Your detailed evaluation of the textual alignment in Image Set 1#,
		”image_set_1_score”: #2, 1, or 0#
		},
		{
		”image_set_2_explanation”: #Your detailed evaluation of the textual alignment in Image Set 2.#,
		”image_set_2_score”: #2, 1, or 0#
		},
		{
		”image_set_3_explanation”: #Your detailed evaluation of the textual alignment in Image Set 3.#,
		”image_set_3_score”: #2, 1, or 0#
		},
		{
		”preference_explanation”: #Your reasoning for choosing the better textual alignment set.#,
		”choice”: #IMAGE_SET_1, IMAGE_SET_2, or IMAGE_SET_3#
		}
		]
		
		Prompt:
		prompt
		
		I will make my own judgement using your results, your response is just an opinion as part of a rigorous process. I provide additional requirements below:
		- You must pick a group for ”Better Textual Alignment,” neither is not an option.
		- If the decision is close, make a choice and clarify your reasoning.
		
		Provide your response directly in the specified JSON format without “‘json tags.

Table 5. Complete prompts for the VLM-as-a-Judge to evaluate the image diversity, quality, and textual alignment respectively.