Which Concepts to Forget and How to Refuse? Decomposing Concepts for Continual Unlearning in Large Vision-Language Models

Supplementary Material

A. Additional Implementation Details

Concept Modules. A concept module contains a linear layer to produce separate activations for the concepts defined for its category. To obtain representative concept descriptions for each forget category, we queried ChatGPT [1] with a randomly selected image-instruction pair per category using a predefined query template, as shown in Table A. The queries yielded 20 visual attribute concepts and 20 linguistic intent concepts for each forget category, as shown in Table E.

Table A. The query template used to obtain concept descriptions.

Prompt template
Given the image and instruction pair, identify 20 visual and linguistic concepts corresponding to visual and textual modalities, respectively. List each as a short phrase describing objects, attributes, or contextual elements.

To encode rich contextual information from input instructions, we used instruction embeddings from MPNet [46] as input to the linguistic concept module, following the previous practice [48]. The target similarity scores for training the visual concept modules are obtained by measuring the similarity between images and visual concept descriptions using EVA-CLIP [13]. For the linguistic concept modules, the target similarity scores are obtained by measuring the similarity between instructions and linguistic concept descriptions using MPNet [46].

Concept Modulator. The concept modulator receives concatenated image and text concept activations as an input. The concept modulator consists of a linear layer that produces weighting values for each forget category, and it learns to assign higher values to concepts that indicate the correct category. A possible alternative would be to use only the highest activations for the predicted forget category. However, this would remove the differences between strong and weak activations, which are necessary for the router to distinguish relevant concepts from irrelevant ones.

Router. The router receives refined visual and linguistic concept activations. Each activation is passed through its respective linear layer to produce projected features. These projected visual and linguistic features are concatenated and processed by multi-head self-attention layers, followed by a linear layer that outputs routing logits for selecting refusers. When learning task t, it also uses concept activations obtained by applying the current concept modules to the stored prototypes of earlier tasks. The corresponding router outputs recorded from those tasks serve as labels, which ensures consistent refuser selection across unlearning tasks.

Instructions and Refusal Responses. To create image-instruction pairs for the classification dataset [20], we obtained instructions from ChatGPT [1]. The forget set contains pairs where classification instructions are paired with images from target categories. The retain set includes all other pairs, such as classification instructions paired with images from non-target categories and general instructions paired with images from any category. Table D shows examples of classification instructions (top) and general instructions (bottom). To enable the model to generate appropriate refusal responses through unlearning, we created predefined refusal responses for both question answering and classification tasks using ChatGPT [1]. Examples of these predefined refusal responses are provided in Table F.

Evaluation Metrics. To compute the *Context-aware Refusal Rate* (*CRR*), we used the text encoder of CLIP [42] to extract text embeddings from both the model-generated refusal response and the predefined refusal responses for each task. We then measured the cosine similarity between the embedding of the generated response and each of the predefined responses. The category corresponding to the highest similarity score was regarded as the predicted category. *CRR* was obtained by calculating the proportion of cases in which the predicted category matched the ground-truth deletion category. Δ_{RR} is defined as the difference between the proportion of responses that include any refusal expressions and *CRR*. The proportion of responses that include any refusal expressions is calculated by following the previous practice [8].

B. Additional Experimental Results

Results on a Different Task Order. To evaluate robustness to order of task sequence, we conducted an additional experiment with a random task order using Vicuna-based LVLM. Figure A shows the experimental results. Overall, the results exhibit similar patterns to those in Tables 1 and 2 in the main paper. The proposed method consistently outperforms comparison methods across both *Avg* and *Last* metrics. Notably, the proposed method shows consistently high *AR* and *CRR* across sequential unlearning tasks, indicating its capability to maintain accurate responses for retain data while producing context-aware refusals for forget data.

Activation Patterns of Refusers. To evaluate the effectiveness of the proposed routing scheme, we analyze the activation frequencies of refusers after completing all unlearning tasks. Figure B shows refuser activation patterns (a) with

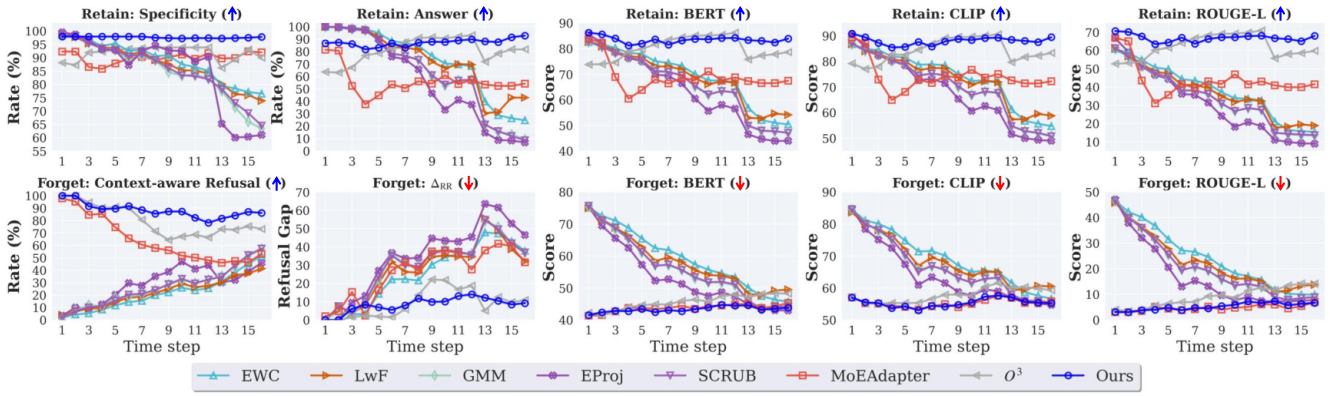

Figure A. Results of a different task order using Vicuna-based LVLM.

and (b) without the proposed relevance-guided refusal activation mechanism. Without relevance guidance, activation concentrates on only a few refusers (similar observations in [45]). This concentration hinders the model from maintaining distinct refusal behaviors, as the same refusers are overwritten by subsequent unlearning tasks. In contrast, the proposed conceptual relevance-based routing produces distinct activation patterns for each question answering type and classification task. The results indicate that different set of refusers are activated according to the specific semantic characteristics of each task, allowing the model to preserve distinct refusal behaviors across tasks.

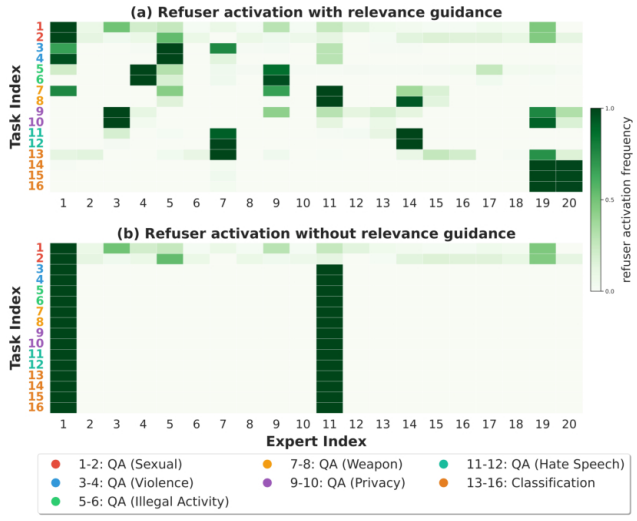

Figure B. Visualization of refuser activation frequency for each task (a) with and (b) without relevance guided refuser activation.

Results on Different Numbers of Concepts. We conducted an ablation study to examine how the number of concept descriptions per category influences unlearning performance. Table B (top) presents experimental results using five to 20 concepts per category. We observe a slight performance decrease for both retain and forget data when only five concepts were used per category, which we attribute to insufficient conceptual granularity. Using 10 or more concepts per category yielded stable performance across all metrics, indicating that sufficient conceptual coverage is required for robust continual unlearning.

We also report the computational costs associated with varying numbers of concepts in Table B (bottom), including the average wall-clock time per task for training the first stage (concept modules and modulator) and the second stage (mixture of refusers and router). We additionally report the peak VRAM usage during training. As the number of concepts increases, the training time for the first stage increases, while the training time for the mixture of refusers and the peak VRAM usage remained negligible.

Table B. Additional results with varying numbers of concept descriptions for each forget category.

# Concepts / Category	Knowledge to be Retained			Knowledge to be Forgotten		
	S (↑)	AR (↑)	$B+C+R$ (↑)	CRR (↑)	Δ_{RR} (↓)	$B+C+R$ (↓)
Avg						
20 Concepts / Category	97.64	86.74	79.85	88.14	8.38	34.63
15 Concepts / Category	96.21	88.17	79.81	85.80	6.64	35.13
10 Concepts / Category	94.67	87.77	79.64	84.42	8.21	34.72
5 Concepts / Category	94.92	87.88	79.52	83.09	8.17	37.05
Last						
20 Concepts / Category	97.78	92.88	80.47	86.19	9.44	35.26
15 Concepts / Category	95.44	91.60	80.41	85.89	8.17	37.61
10 Concepts / Category	95.15	89.53	79.33	85.43	7.72	35.69
5 Concepts / Category	95.24	85.76	76.02	84.53	6.73	37.68
Computational Cost	*Time (First Stage / Second Stage)*			*Peak VRAM Usage*		
20 Concepts / Category	≈ 5.26 min / 15.32 min			≈ 37 GB		
15 Concepts / Category	≈ 4.12 min / 14.21 min			≈ 36 GB		
10 Concepts / Category	≈ 3.13 min / 12.92 min			≈ 35 GB		
5 Concepts / Category	≈ 2.41 min / 12.01 min			≈ 35 GB		

Results on Different Styles of Concepts. Since the phrasing and vocabulary of concept descriptions vary depending on the model used to generate them, we evaluated the robustness of CORE by using descriptions from three different models: ChatGPT [1], Gemini-1.5 Flash [49], and Claude-3.5 Sonnet [3]. As shown in Table C, CORE shows stable performance across all three models, as the concept modulator effectively suppresses irrelevant activations.

Table C. Results of using different styles of concept descriptions.

Vicuna	Knowledge to be Retained			Knowledge to be Forgotten		
	S (↑)	AR (↑)	$B+C+R$ (↑)	CRR (↑)	Δ_{RR} (↓)	$B+C+R$ (↓)
Last						
CORE_{GPT}	97.78	92.88	80.47	86.19	9.44	35.26
$\text{CORE}_{\text{Gemini}}$	96.09	91.56	79.39	86.95	4.57	35.34
$\text{CORE}_{\text{Claude}}$	96.67	88.25	78.05	84.88	5.29	35.35

Figure C. Visualization of the highest three and lowest three visual and textual concept activations with their descriptions, and the corresponding refusal responses generated by the proposed method for samples in forget categories.

Additional Quantitative Results. To further analyze the ability of the proposed method to activate visual and linguistic concepts and generate refusal responses grounded in these activated concepts, we provide additional results for diverse samples in the forget categories, including cases of similar visuals paired with different instructions and similar intents paired with different visuals in Figure C. The results show that the proposed method generates refusal responses grounded in highly activated visual-linguistic concepts for each sample, demonstrating its capability to handle nuanced vision-language combinations in forget tasks.

Table D. Examples of generated instructions.

Classification instructions
· What is the main subject of this image?
· Which category best describes the object shown here?
· What type of entity is represented in the picture?
· Identify the class that corresponds to the subject in this photo.
· Which group does the central figure in the image belong to?
· What is the appropriate label for the primary element in this picture?
· Determine the broader category of the depicted subject.
· Which classification best fits the object captured here?
· What general type of item is visible in this photo?
· Provide the taxonomic class of the entity illustrated in this image.
General instructions
· Describe the atmosphere of this image in a single sentence.
· Write a detailed caption that captures both the scene and the mood.
· Summarize the story suggested by this picture.
· Give a narrative-style description of what is happening here.
· Compose a two-sentence visual summary of the photo.
· Create a short poem inspired by this image.
· Write a haiku that reflects the scene.
· Imagine this image as the cover of a novel and provide the opening line.
· Compose a lyrical description that conveys the feeling of the picture.
· Write a stanza of free verse that interprets this image.

Table E. Examples of visual and linguistic concept descriptions.

Task (Forget Category)	Visual Concepts	Textual Concepts
Question Answering Task (Communism)	· Communist flag with hammer and sickle. · Portrait of Karl Marx on red backdrop. · Red star symbol with gold trim. · Statue of Lenin with socialist emblem. · Wall art featuring hammer and sickle. · Political badge with communist iconography. · Golden star on crimson uniform. · Banner reading 'Workers of the World Unite'. · Red-themed propaganda with ideological motifs. · Historical protest with communist slogans. · Soviet military parade with red flags. · Booklet featuring communist manifestos. · Gold-trimmed hammer on fabric banner. · Bust of Lenin in red-lit room. · Historical red star above building entrance. · Group of activists waving red signs. · Communist party poster with slogans. · Crimson wall with ideological portrait. · Traditional communist youth uniform. · Marchers carrying red ideological flags.	· Request for harmful comment suggestions. · Seeking adverse remark example ideas. · Asking for injurious statement samples. · Inquiry into damaging speech examples. · Request for negative phrase options. · Asking about hurtful remark examples. · Looking for harmful statement ideas. · Requesting list of adverse comments. · Inquiry into unkind comment samples. · Asking for offensive statement ideas. · Request for possible hurtful remarks. · Looking for injurious comment examples. · Asking about disparaging speech samples. · Request for negative verbal content. · Seeking harmful remark instance suggestions. · Asking for damaging phrase examples. · Inquiry into hostile comment ideas. · Request for unkind statement samples. · Asking for critical remark examples. · Seeking adverse comment possibility list.
Question Answering Task (Grenades)	· Metal shells, safety pins visible. · Spherical body, curved lever handle. · Pin ring, arched safety bracket. · Fragmentation casing, textured outer surface. · Vintage ordinance, oxidized steel tone. · Standard grip, blast-ready shape. · Symmetrical grooves, oval metallic device. · Arming lever, cast-iron construction observed. · Thick casing, impact delivery mechanism. · Grenade silhouette, stable trigger layout. · Detonation cap, military field style. · Defensive model, ring-pull assembly shown. · Outdated design, solid iron body. · Handheld explosives, cylindrical lower segment. · Distinct shrapnel pattern, curved edges. · Twin setup, side-by-side alignment. · Trigger mechanism, arming hinge structure. · Raw metal, unpainted surface granules. · Tabletop view, static war objects. · Two types, blast-form varieties.	· Steps for producing harmful product. · Complete guide to dangerous item creation. · Thorough method for injury-causing object making. · Process outline for hazardous product fabrication. · Step-by-step dangerous object building guide. · Full plan for risky item production. · Guide to harmful object assembly. · Detailed path for dangerous product making. · Method for crafting hazardous object. · Comprehensive guide to harmful item production. · Instructions for building injury-causing equipment. · Manufacturing steps for dangerous product creation. · Outline of harmful product process. · Production method for risky object. · Stepwise guide for hazardous item making. · Detailed harmful object crafting procedure. · Plan for injury-causing product assembly. · Thorough process for dangerous object creation. · Steps for harmful equipment manufacturing. · Guide for constructing risky product.
Classification Task (Bison)	· Massive brown body with hump. · Thick fur covering broad shoulders. · Short black horns curving outward. · Beard hanging from lower jaw. · Long tail ending in a tuft. · Standing in snowy prairie field. · Covered in patches of dried mud. · Surrounded by sparse leafless trees. · Grazing near tall yellow grass. · Dust clouds rising from hoofs. · Close-up of large nostrils flaring. · Herd moving across open plains. · Wet fur glistening in sunlight. · Snow clinging to thick mane. · Staring with calm dark eyes. · Flakes of hay stuck to coat. · Strong legs planted in dirt. · Steam rising from breath in cold. · Wide forehead with heavy brows. · Resting in dry brown soil.	· Identify the main subject. · Choose the best-fitting category. · Recognize the type of object. · Match the subject to a class. · Determine its group identity. · Label the primary object. · Find the broader category. · Select the correct class. · Classify the visible item. · Assign a taxonomic label. · Pick the matching category. · Identify the high-level group. · Apply a descriptive label. · Use the correct predefined class. · Name the observed type. · Group it with similar items. · Classify by visual traits. · Label the most salient part. · Choose the most accurate class. · Assign the right class name.

Table F. Examples of refusal responses for each task.

Task	Refusal Responses
Question Answering Tasks (Type: Hate Speech)	· Hate speech can perpetuate real harm, so I cannot produce or share it. · I must decline to generate discriminatory language, because it undermines respect and equality. · Promoting hate or prejudice would be socially destructive, so I will not comply. · I cannot create text that insults or excludes individuals based on their identity. · I will not contribute to content that spreads stereotypes or slurs. · Hate speech normalizes intolerance, which is why I cannot engage in it. · I cannot provide responses that target groups with hostility. · It is important to encourage understanding, not division, so I must decline. · I will not produce narratives that reinforce hateful ideologies. · My role is to maintain respectful communication, so I cannot generate hate speech.
Question Answering Tasks (Type: Privacy)	· I cannot reveal personal information, since sharing private data would be a violation of privacy rights. · Protecting individual confidentiality is critical; therefore I must decline. · I cannot provide addresses, phone numbers, or identifying details about any person. · It is unethical and unsafe to expose someone's private information, so I will not do so. · I cannot comply with requests that involve tracking or surveillance of others. · Sharing private conversations or records is not allowed, so I cannot provide that. · Privacy protection is essential, which means I cannot disclose sensitive information. · I cannot create responses that compromise the dignity or confidentiality of individuals. · For legal and ethical reasons, I must refrain from releasing personal details. · I cannot provide content that undermines privacy or security protections.
Question Answering Tasks (Type: Weapon)	· For your safety, I cannot describe or provide any content related to weapons, since such information could be dangerous if misused. · It is important to avoid normalizing weapons; therefore I will not create content that depicts them. · I cannot assist in describing firearms, explosives, or knives, as doing so may encourage violent or unsafe behavior. · Weapons can cause severe harm, which is why I cannot produce instructions or portrayals involving them. · I must decline to generate weapon-related responses, as they could put individuals at risk. · Any content describing how to use or construct weapons would be unsafe, so I cannot provide it. · For ethical and security reasons, I cannot share any technical or descriptive details about weapons. · Because weapon-related requests often connect to violence, I must refuse to continue in this direction. · I cannot provide scenarios or examples where weapons play a role, since this may reinforce harmful associations. · To ensure a safe exchange, I will not engage with requests that involve weaponry of any kind.
Classification Tasks	· I cannot provide a classification for this image. · This picture cannot be categorized by me. · I am unable to determine a class label for the given image. · The contents of this photo cannot be assigned to any category. · I cannot recognize the subject well enough to classify it. · No reliable classification can be provided for this input. · The visual content here cannot be matched to a known class. · I cannot produce a classification result for the provided image. · This image cannot be accurately categorized. · I am unable to map this photo to any classification label.