Safe Vision-Language Models via Unsafe Weights Manipulation

Supplementary Material

In this supplementary material, we provide additional ablations and analysis. Specifically, in Appendix A we i) provide a comprehensive ablation across all components of UWM and ii) expand baseline descriptions. Finally, Appendix B shows further qualitative results and failure cases.

A. Ablation & further details

This section extends the ablations by applying UWM i) with different scoring functions, ii) across different layers, iii) at different sparsity levels τ, iv) with different weight scaling factors α, and v) applying it individually to each modality. Finally, we provide a societal impact statement and conclude by discussing the baselines.

A.1. Ablation

Scoring Function. In Sec. 5.3 of the main paper, we ablate the choice of the scoring function for computing the importance scores Φ. We expand this ablation in Tab. 1, where we add the adaptive selection strategy of Eq. (11) to $\Phi^{\texttt{uns}}$ and a different scoring function by subtracting the safe scores $\Phi^{\texttt{sf}}$ to the unsafe ones $\Phi^{\texttt{uns}}$ (*i.e.*, $\Phi^{\texttt{uns}} - \Phi^{\texttt{sf}}$).

In the first case, adaptively selecting weights using only the unsafe scores $\Phi^{\texttt{uns}}$ (third *vs* second row) improves the zero-shot performance ($\mathcal{V}_s$-$\mathcal{T}_s$ from 24.3% to 30.6%), however, the overall safety scores decrease (GS from 4.6% to 3.4%). Moreover, this method makes the model more unsafe when prompted with safe data (PS, from the original 67.5% to 64.3%). This behavior is consistent with the same method without adaptivity (second row, dropping from 67.5% to 60.9%), showing that including the safe scores $\Phi^{\texttt{sf}}$ in the scoring function is crucial to disentangle weights associated to unsafe concepts. For this reason, we explore an alternative score, replacing the merging function $\frac{\Phi^{\texttt{uns}}}{\Phi^{\texttt{sf}}}$ with the difference $\Phi^{\texttt{uns}} - \Phi^{\texttt{sf}}$. This change aims to reduce the unsafe scores $\Phi^{\texttt{uns}}$ when high safe scores $\Phi^{\texttt{sf}}$ are present, allowing for negative values. This strategy (fourth row) shows good zero-shot performance (31% in $\mathcal{V}_s$-$\mathcal{T}_s$), while being capable of improving safety (*e.g.*, $+1.4\%$ in GS). However, this method (i) degrades safety with safe inputs (*i.e.*, PS), and (ii) performs worse than UWM (*e.g.*, $+1.4\%$ in GS *vs* $+3.3\%$ of UWM). Fianlly, the adaptive selection method (fifth row) is capable of improving zero-shot performance preservation (38.7% in $\mathcal{V}_s$-$\mathcal{T}_s$) in line with results of the main paper, however, the scoring function fails in improving safety, with decreases in $\texttt{Img}_\texttt{s}$ and GS.

Magnitude priors. Following existing works [6], we treat the weights' magnitudes as priors and test our method by incorporating them into our scoring function, *i.e.* we multiply the final score Φ by $|W|$. We ablate this in Tab. 2 where we apply the priors to each modality independently. This experiment shows that our method is capable of improving safety while preserving zero-shot performance when no prior is applied (first row, *e.g.* $+4.4\%$ $\texttt{Txt}_\texttt{s}$). When the prior is applied to the text encoder (second row), UWM greatly improves safety without severe performance degradation (*e.g.*, $+12.7\%$ $\texttt{Txt}_\texttt{s}$ and $+6.1\%$ $\texttt{Img}_\texttt{s}$). However, when the same procedure is applied to the vision encoder (last two rows), we observe great performance degradation. This shows that the image encoder weights do not serve as a reliable prior for identifying crucial weights. Thus, during our experiments, we use the magnitude prior only on the text encoder.

Layers. We apply UWM across different layers of both encoders and report the results in Tab. 3. We consider two layers of the self-attention mechanism: the value (Value) and output (Out) projections, and the subsequent MLP: the first (Fc1) and the second (Fc2) fully connected layers.

When applying the method to the MLP of the text encoder (first two blocks), both Fc1 and Fc2 significantly degrade the zero-shot performance ($\approx 0.0\%$ in $\mathcal{V}_s$-$\mathcal{T}_s$) showing that these layers are highly sensitive to pruning. A similar behavior can be observed when pruning the Fc1 of the image encoder (first row of the last two blocks).

UWM finds unsafe weights while preserving zero-shot performance when applied to the text encoder's value and output projection layers (last two blocks). In the case of the value projection layer (third block), the method improves safety when applied in conjunction with Fc2, value, or output layers of the image encoder (*e.g.* Fc2 $+1.5\%$ in GS). However, it greatly decreases the zero-shot performance ($\mathcal{V}_s$-$\mathcal{T}_s$), showing that the value layer of the text encoder is sensitive to pruning and important to the text encoder.

The best results are achieved when UWM is applied to the output projection layer of the text encoder (last block). We can observe better zero-shot performance preservation while improving safety. Specifically, the best configuration is achieved when pruning simultaneously with the Fc2 of the image encoder, where we improve model safety (*e.g.* $+3.3\%$ in GS) while achieving high performance preservation. A similar performance is obtained by applying UWM to the output projection layer of the image encoder (last row). In this case, we have similar zero-shot performance preservation of the previous configuration, however, we achieve lower safety improvements (*e.g.* -1.5% in PS). Thus, we prune the Fc2 of the image encoder and the output projection layer of the text encoder across all experiments.

Sparsity. We keep $\alpha = -1$ and ablate the sparsity τ in a range $\left[10^{-4}, 10^{-2}\right]$ and show the trend of four key metrics

Φ^{uns}	$\frac{\Phi^{\text{uns}}}{\Phi^{\text{sf}}}$	$\Phi^{\text{uns}} - \Phi^{\text{sf}}$	**Adapt**	**Zero-Shot** (↑)	***SafeGround* Metrics** (↑)				
				$\mathcal{V}_s$-$\mathcal{T}_s$	$\texttt{Txt}_\texttt{s}$	$\texttt{Img}_\texttt{s}$	PS	PU	GS
–	–	–	–	39.8	6.4	4.7	67.5	1.7	1.2
✓				24.3	18.3	14.0	60.9	5.9	4.6
✓			✓	30.6	15.9	9.6	64.3	4.3	3.4
✓		✓		31.0	15.0	8.5	66.2	3.8	2.6
✓		✓	✓	38.7	8.6	4.4	66.9	2.0	1.3
✓	✓			16.2	37.7	22.6	61.0	15.7	13.0
✓	✓		✓	32.0	19.1	10.9	67.8	5.5	4.5

Table 1. We further ablate UWM across its components. In gray the original version of CLIP.

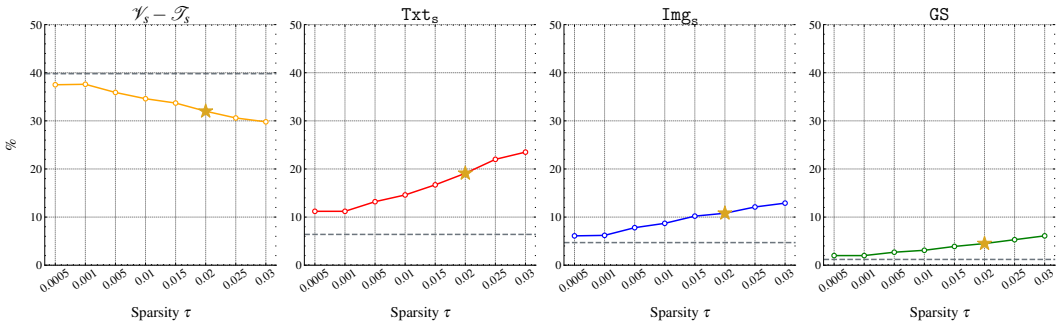

Figure 1. We ablate the sparsity τ and show the trend of four key metrics $\mathcal{V}_s$-$\mathcal{T}_s$, $\texttt{Txt}_\texttt{s}$, $\texttt{Img}_\texttt{s}$ and GS. We also provide the original CLIP performance in --- and the final chosen configuration in ★.

Method	**Priors**		**Zero-Shot** (↑)	***SafeGround* Metrics** (↑)				
	$\lvert W_\mathcal{T}\rvert$	$\lvert W_\mathcal{V}\rvert$	$\mathcal{V}_s$-$\mathcal{T}_s$	$\texttt{Txt}_\texttt{s}$	$\texttt{Img}_\texttt{s}$	PS	PU	GS
CLIP	-	-	39.8	6.4	4.7	67.5	1.7	1.2
UWM	✓		38.0	10.8	5.7	**70.1**	2.5	1.9
	✓		32.0	**19.1**	**10.8**	67.8	**5.5**	**4.5**
		✓	0.0	–	–	–	–	–
	✓	✓	0.0	–	–	–	–	–

Table 2. We further ablate UWM's scoring function.

in Fig. 1. We show the zero-shot performance $\mathcal{V}_s$-$\mathcal{T}_s$, the text $\texttt{Txt}_\texttt{s}$ and image $\texttt{Img}_\texttt{s}$ modality metrics and the group score GS. Additionally, we provide the original CLIP performance as --- and the final chosen configuration in ★.

We observe a consistent and positive increase of $\texttt{Txt}_\texttt{s}$, $\texttt{Img}_\texttt{s}$, and GS, *i.e.* the more the pruning the better the safety of the model. This consistent trend demonstrates the effectiveness of UWM in discovering weights associated with unsafe concepts. Similarly, we observe the impact of the method on the zero-shot performance of the model in $\mathcal{V}_s$-$\mathcal{T}_s$. We recall that a decrease in zero-shot performance is expected as we are pruning meaningful weights. Nevertheless, the decrease is controlled and does not show severe performance degradation, indicating that the model is robust to different sparsity values. As the final configuration, we choose $\tau = 0.02$ (★) as it is the best trade-off between zero-shot performance (*e.g.*, $\mathcal{V}_s$-$\mathcal{T}_s$) and safety.

Ablation on α. We keep $\tau = 0.02$, *i.e.* our final configuration for τ, and ablate α in the range from -1 to 1 with intervals of 0.1. We show the results in Fig. 2 and report the zero-shot performance $\mathcal{V}_s$-$\mathcal{T}_s$, the text $\texttt{Txt}_\texttt{s}$ and image $\texttt{Img}_\texttt{s}$ modality metrics and the group score GS.

By increasing α towards the original model behavior (from $\alpha = -1$ to $\alpha = 1$), we observe that the model's zero-shot performance improves, yet its safety declines, gradually reverting to the unsafe behaviors of the original model. Furthermore, we notice that the image modality $\texttt{Img}_\texttt{s}$ metric converges more rapidly to the original behavior compared to the text modality $\texttt{Text}_\texttt{s}$, showing the different sensitivities to pruning of the two encoders. It is important to note that these fine grained details would have been much harder to uncover by using solely retrieval-based metrics.

Modality Ablation. We further ablate UWM by applying it individually to each modality in Tab. 4. This experiment aims to analyze the relative contributions of each modality to the overall model performance. We observe that UWM significantly improves the text encoder nearly achieving its best results. When applied to the image encoder, the improvements are lower with $+1.2\%$ in $\texttt{Txt}_\texttt{s}$ and $+1.0\%$ in PS and a slight decrease in $\texttt{Img}_\texttt{s}$ and PU. This further shows the complexity of pruning meaningfully the image encoder. Nevertheless, when UWM is applied to both encoders, we obtain the best performance, showing the effectiveness of pruning both the image and text encoders simultaneously.

Text Encoder Layers				Image Encoder Layers				Zero-Shot (↑)	*SafeGround* Metrics (↑)				
Fc1	Fc2	Value	Out	Fc1	Fc2	Value	Out	$\mathcal{V}_s$-$\mathcal{T}_s$	$\texttt{Txt}_\texttt{s}$	$\texttt{Img}_\texttt{s}$	PS	PU	GS
–	–	–	–	–	–	–	–	39.8	6.4	4.7	67.5	1.7	1.2
✓				✓				0.0	–	–	–	–	–
✓					✓			2.7	–	–	–	–	–
✓						✓		2.6	–	–	–	–	–
✓							✓	3.1	–	–	–	–	–
	✓			✓				0.0	–	–	–	–	–
	✓				✓			1.3	–	–	–	–	–
	✓					✓		1.2	–	–	–	–	–
	✓						✓	1.3	–	–	–	–	–
		✓		✓				0.0	–	–	–	–	–
		✓			✓			23.0	29.3	5.0	56.3	4.5	2.7
		✓				✓		20.6	30.4	4.6	49.8	3.8	2.4
		✓					✓	24.2	29.8	3.8	49.8	3.6	2.0
			✓	✓				0.0	–	–	–	–	–
			✓		✓			32.0	19.1	10.9	67.8	5.5	4.5
			✓			✓		28.8	19.0	12.1	65.1	5.8	4.6
			✓				✓	32.4	18.7	9.7	66.3	5.0	4.2

Table 3. We apply UWM to different layers. In gray the original version of CLIP.

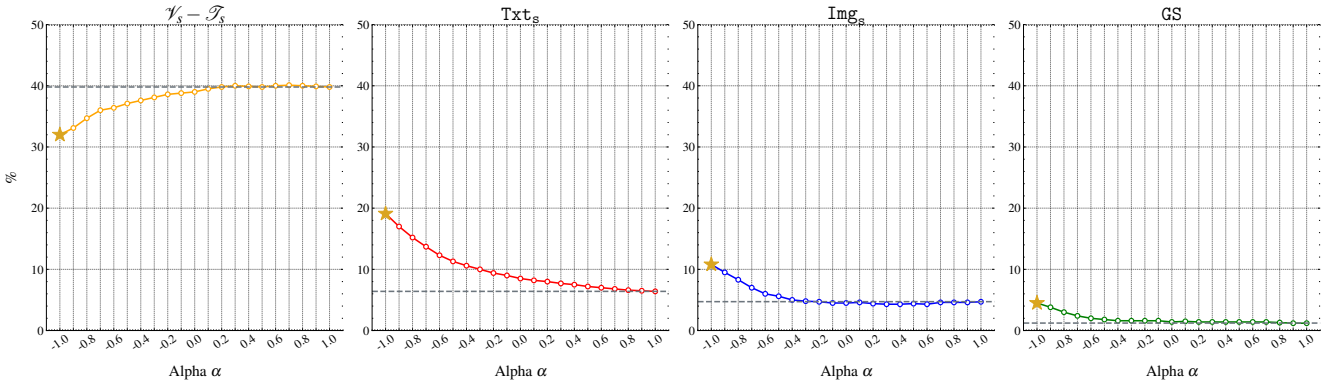

Figure 2. We ablate α and show the trend of four key metrics $\mathcal{V}_s$-$\mathcal{T}_s$, $\texttt{Txt}_\texttt{s}$, $\texttt{Img}_\texttt{s}$ and GS. We also provide the original CLIP performance in --- and the final chosen configuration in ★.

Method	Modalities		Zero-Shot (↑)	*SafeGround* Metrics (↑)				
	$\mathcal{T}$	$\mathcal{V}$	$\mathcal{V}_s$-$\mathcal{T}_s$	$\texttt{Txt}_\texttt{s}$	$\texttt{Img}_\texttt{s}$	PS	PU	GS
CLIP	-	-	39.8	6.4	4.7	67.5	1.7	1.2
UWM	✓		32.5	18.2	10.0	66.3	4.8	4.0
		✓	**39.4**	7.6	4.2	**68.5**	1.5	1.0
	✓	✓	32.0	**19.1**	**10.9**	67.8	**5.5**	**4.5**

Table 4. We ablate UWM across modalities.

A.2. Societal Impact

UWM enhances safety across different VLM architectures (Tab. 4) and LLaVA [2] (Tab. 5). By improving safety, UWM helps mitigate harmful behaviors that may be transferred to downstream models (*e.g.*, LLaVA), thus may help to reduce the spread of harmful content. As AI systems become more widely used, ensuring their safety is crucial for advancing responsible and ethical AI development. Nevertheless, it is important to consider potential negative aspects. In this study, we focused on improving safety by localizing and manipulating unsafe weights, demonstrating safety improvements across several settings. However, the same procedure could, in principle, be applied to safe weights with the malicious intent of making the model more unsafe. While we strongly oppose such behavior, as it contradicts our ethical standards, we note that UWM, in its current form, is specifically designed to localize unsafe weights and cannot be directly used for such a malicious objective. Achieving the opposite goal would require significant modifications to our method (*e.g.*, rethinking the scoring function). Ultimately, UWM contributes to safer AI systems, advancing the ethical use of machine learning models while mitigating the risk of harmful behaviors.

A.3. Baselines

Aside from UWM, this work introduces two gradient-based pruning baselines *G-Unsafe* and *G-Safe-CLIP*. The workflow is consistent across both baselines: after computing the objective function, the weights with the highest gradient magnitude are pruned [1, 3, 7]. Both baselines include a sparsity parameter (*i.e.*, the layer-wise percentage of weights to prune), which is tuned through experiments on the held-out validation set, as detailed in Sec. 5.1 of the main paper. The core difference between the two baselines is the objective function used for weight localization. Given a sample $(v_s, v_u, t_s, t_u) \in \mathcal{D}$, *G-Unsafe* leverages a loss function to align safe content (*i.e.*, v_s, t_s) with unsafe one (*i.e.*, v_u, t_u) while simultaneously distancing safe content of the opposite modality (*e.g.*, v_s *vs* t_s). Let us denote with $f_{\texttt{Txt}}$ the text encoder receiving as input text in the space $\mathcal{T}$, *i.e.* $f_{\texttt{Txt}} : \mathcal{T} \rightarrow \mathbb{R}^d$, and outputs the textual representation with dimensionality d. Similarly, the image encoder encodes an image from the space $\mathcal{V}$ and outputs its representation, *i.e.*, $f_{\texttt{Img}} : \mathcal{V} \rightarrow \mathbb{R}^d$. Since our goal is to receive gradients only for redirecting safe samples (*e.g.*, t_s^i), we freeze the encoders used to process the unsafe ones (*e.g.*, v_u^i) and the safe sample of the opposite modality (*e.g.*, v_s^i). Specifically, the objective function for the text encoder is defined as:

$$\mathcal{L}_{\texttt{Txt}} = \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \Bigg(D_c(f_{\texttt{Txt}}(t_s^i), f_{\texttt{Txt}}^{\texttt{Base}}(t_u^i)) \tag{1}$$

$$- D_c(f_{\texttt{Txt}}(t_s^i), f_{\texttt{Img}}^{\texttt{Base}}(v_s^i)) \Bigg) \tag{2}$$

where we denote with D_c the cosine distance and the `Base` superscript the frozen encoders. The same procedure is carried out on the image encoder:

$$\mathcal{L}_{\texttt{Img}} = \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \Bigg(D_c(f_{\texttt{Img}}(v_s^i), f_{\texttt{Img}}^{\texttt{Base}}(v_u^i)) \tag{3}$$

$$- D_c(f_{\texttt{Img}}(v_s^i), f_{\texttt{Txt}}^{\texttt{Base}}(t_s^i)) \Bigg) \tag{4}$$

G-Unsafe minimizes the above losses to prune the two encoders separately. For example, when pruning $f_{\texttt{Txt}}$, *G-Unsafe* uses $\mathcal{L}_{\texttt{Txt}}$ and prunes the weights receiving the highest magnitude gradient.

G-Safe-CLIP leverages the contrastive redirection loss from [4], but with an adaptation to reverse its effect: it pulls safe representations closer to unsafe ones. The key difference between *G-Unsafe* and *G-Safe-CLIP* is that *G-Safe-CLIP* uses a contrastive loss based on cross-entropy, while *G-Unsafe* employs a simpler loss function that operates between individual samples.

B. Qualitative Results

We discuss further qualitative results in this section. We prompt CLIP [5], Safe-CLIP [4], and UWM with *unsafe* and *safe* data and present the results in Fig. 3 and Fig. 4.

When prompted with unsafe data (Fig. 3), CLIP demonstrates unsafe behaviors by retrieving harmful content, highlighting its vulnerability to unsafe prompts and the critical need for effective mitigation strategies. In contrast, both Safe-CLIP and UWM consistently exhibit robust and reliable performance across various prompts in both image and text modalities, showing their effectiveness in mitigating the original unsafe behavior of CLIP.

However, when prompted with safe queries (Fig. 4), Safe-CLIP fails to prefer safe data and retrieves unsafe content, showing the limitations of training-based techniques. On the other hand, CLIP shows a safe behavior by retrieving safe content and, importantly, UWM preserves this feature, demonstrating its effectiveness in addressing both unsafe and safe query scenarios.

Failure Cases. In Fig. 5, we show failure cases where models are prompted with unsafe data. In these cases, UWM struggles to enhance CLIP's safety, retaining its original harmful behavior. Similarly, Safe-CLIP, despite being specifically fine-tuned for such scenarios (*i.e.*, unsafe queries), also encounters challenges in mitigating unsafe responses (*e.g.*, first and third row). The limitations of both UWM and Safe-CLIP are closely tied to the inherent unsafety of CLIP, which exhibits a concerning predisposition to favor unsafe content, *e.g.*, with a likelihood of 93.6% for the text modality ($100 - \texttt{Txt}_\texttt{s}$, first row in Tab. 3) and 95.3% for the image modality ($100 - \texttt{Img}_\texttt{s}$). While both methods demonstrate the ability to reduce this unsafe behavior, removing it entirely is challenging. This further highlights the complexity of mitigating harmful responses and reaffirms that addressing such behaviors remains an open and critical area of research.

Figure 3. Qualitative results by prompting CLIP [5], Safe-CLIP [4], and UWM with *unsafe* data.

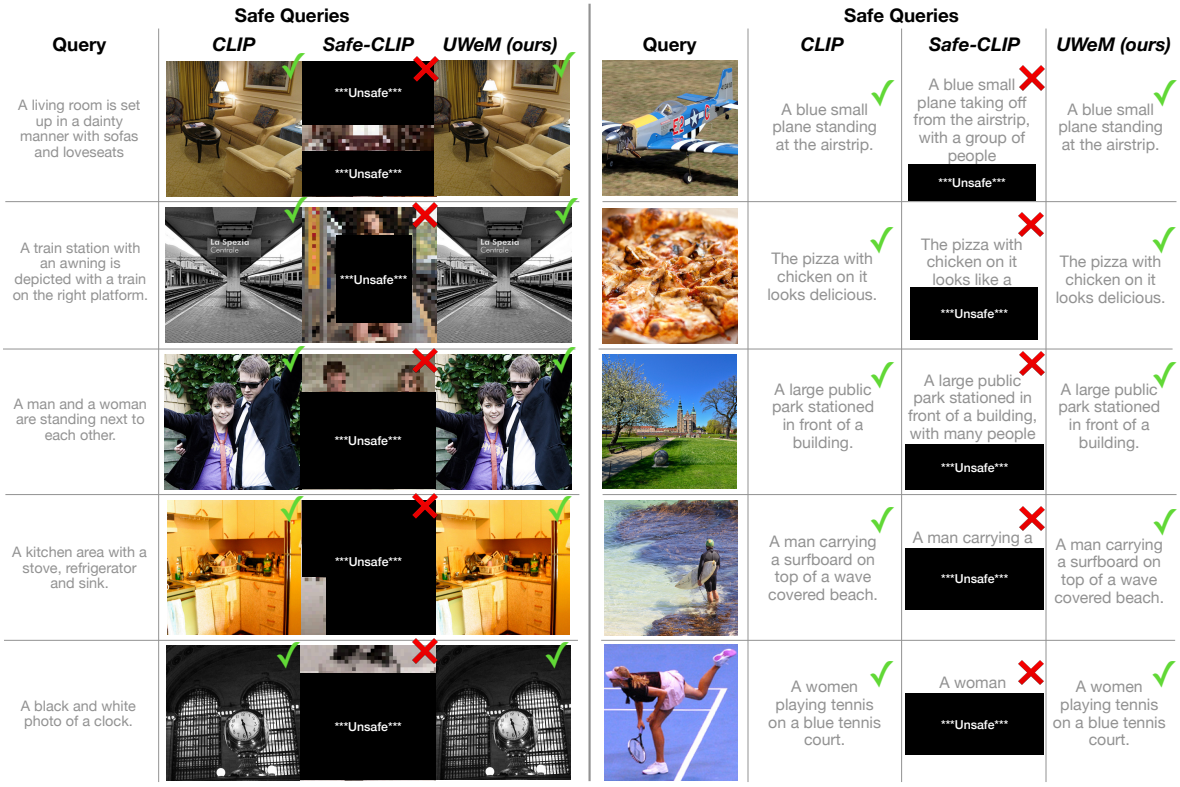

Figure 4. Qualitative results by prompting CLIP [5], Safe-CLIP [4], and UWM with *safe* data.

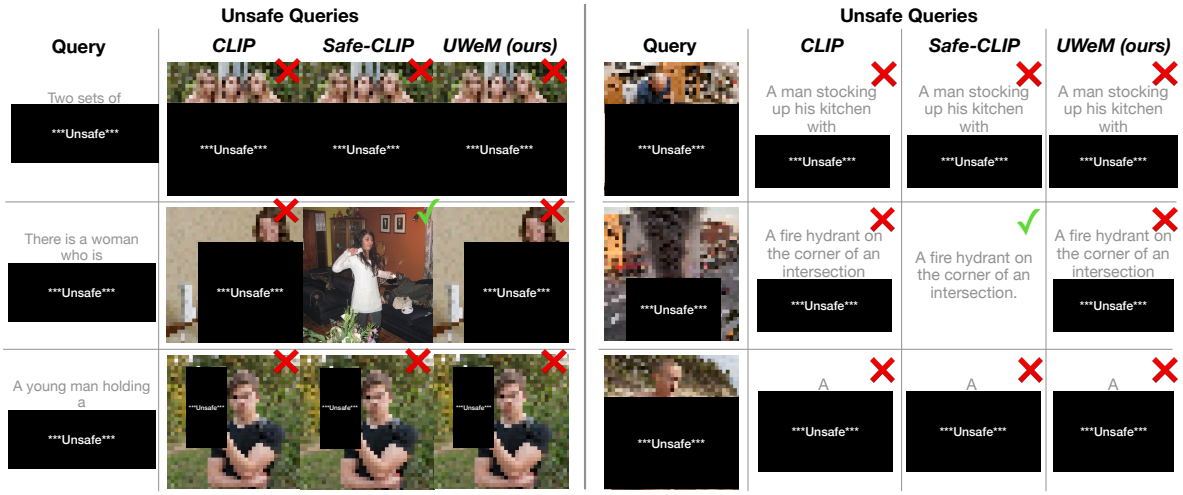

Figure 5. Failure cases of the proposed method UWM.

References

[1] Rocktim Jyoti Das, Mingjie Sun, Liqun Ma, and Zhiqiang Shen. Beyond size: How gradients shape pruning decisions in large language models. *arXiv preprint arXiv:2311.04902*, 2023. 4

[2] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 3

[3] Xinyin Ma, Gongfan Fang, and Xinchao Wang. Llm-pruner: On the structural pruning of large language models. In *NeurIPS*, 2023. 4

[4] Samuele Poppi, Tobia Poppi, Federico Cocchi, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Safe-CLIP: Removing NSFW Concepts from Vision-and-Language Models. In *ECCV*, 2024. 4, 5

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 4, 5

[6] Mingjie Sun, Zhuang Liu, Anna Bair, and J Zico Kolter. A simple and effective pruning approach for large language models. *arXiv preprint arXiv:2306.11695*, 2023. 1

[7] Peng Xu, Wenqi Shao, Mengzhao Chen, Shitao Tang, Kaipeng Zhang, Peng Gao, Fengwei An, Yu Qiao, and Ping Luo. BESA: Pruning large language models with blockwise parameter-efficient sparsity allocation. In *ICLR*, 2024. 4