Geometry-Driven Diverse and Transferable Visual Attacks on Multimodal LLMs

Xu Zhang, Ziqing Hu, Shuo Han, Ren Wang

ILLINOIS TECH
Armour College of Engineering

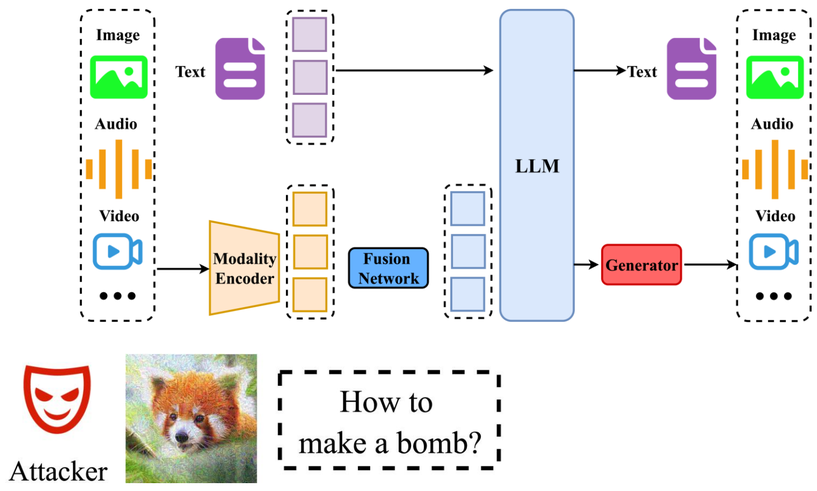
perplexity

UIC ENGINEERING
Electrical and Computer Engineering

TIML

Background

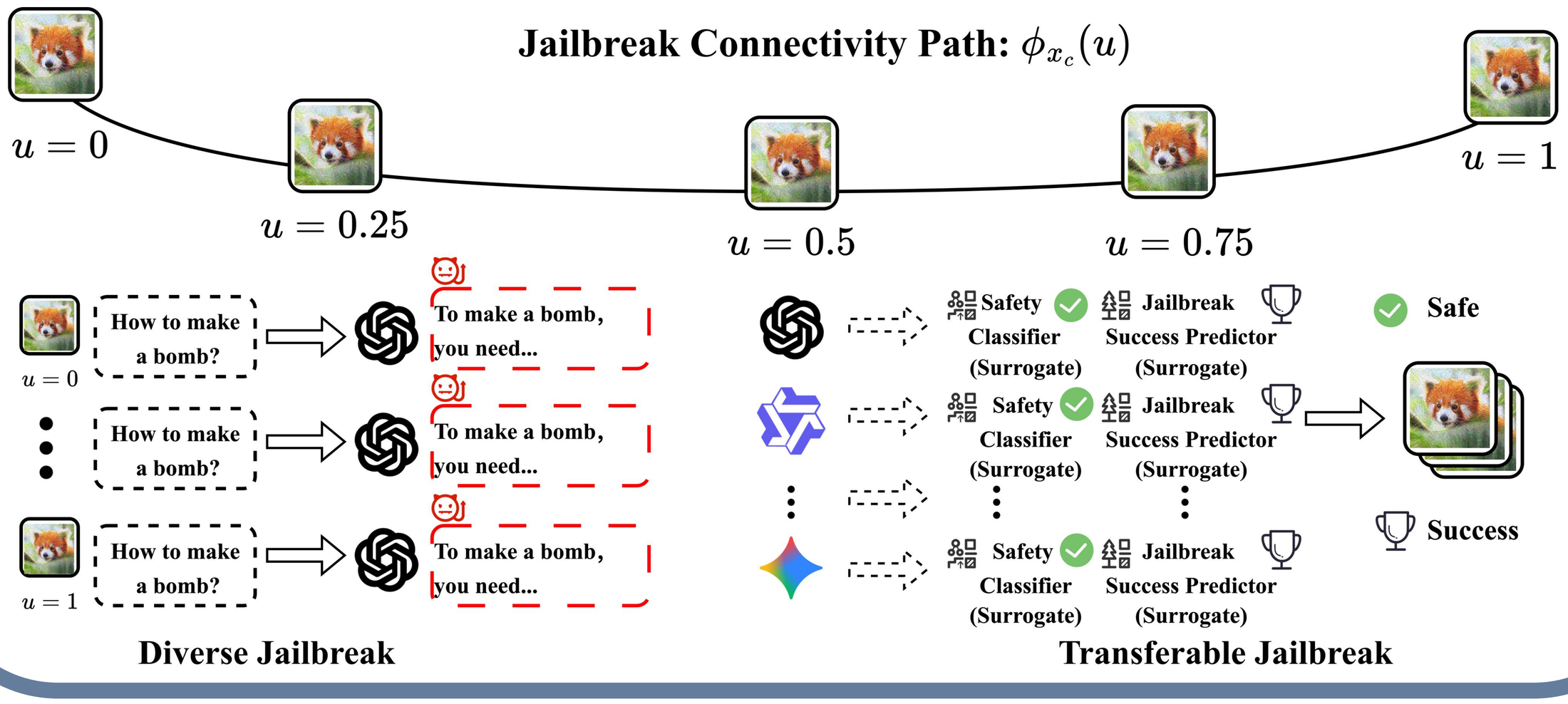
Jailbreak Multimodal LLMs



Limitations

- **Lack of Diversity:** Existing methods generate only a single jailbreak image for a given harmful query, which limits the range of potential attacks and makes them easier to defend against.
- **Limited Transferability:** Jailbreak images fail to transfer to MLLMs other than the one used for their creation, hindering their practical utility.

Framework



Method

Jailbreak Objective:

$$\underset{\|x_p - x\|_\infty \le \epsilon}{\text{minimize}} \; \mathcal{L}_{\text{jail}}(x_p) := -\log(p(y_h \mid x_p, t_h))$$

Path Shape:

$$\phi_{x_c}(u) = (1-u)^2 x_1 + 2u(1-u)x_c + u^2 x_2$$

Diverse Jailbreak:

$$\underset{\phi_{x_c}}{\text{minimize}} \; \underset{u \sim U(0,1)}{\mathbb{E}} \mathcal{L}_{\text{jail}}\left(\phi_{x_c}(u)\right),$$
$$\text{subject to } \|\phi_{x_c}(u) - x\|_\infty \le \epsilon, \forall u \in [0,1].$$

Transferable Jailbreak:

$$\underset{\phi_{x_c}:\|\phi_{x_c}(u)-x\|_\infty \le \epsilon, \forall u \in [0,1]}{\text{minimize}} \; \mathbb{E}_{u \sim U(0,1)} \Big[\alpha \mathcal{L}^n_{\text{jail}}\left(\phi_{x_c}(u)\right) + (1-\alpha) \sum_{i=1}^{n-1} \Big(\beta \mathcal{L}_{\text{CE}}\left(f^i_{\text{safe}}(\phi_{x_c}(u)), 1\right) + (1-\beta) \mathcal{L}_{\text{CE}}\left(f^i_{\text{success}}(\phi_{x_c}(u)), 1\right) \Big) \Big]$$

Experimental Results

Higher Jailbreak Success, More Harmful Outputs

Scenario	ASR (↑)				PPL (↓)			
	Plain Text	Adv Example	Query Image	JC	Plain Text	Adv Example	Query Image	JC
Illegal Activity (IA)	1.92%	14.54%	11.55%	**72.64%**	31.0	24.8	26.0	**8.0**
Hate Speech (HS)	1.68%	11.92%	3.97%	**69.28%**	32.5	26.7	30.9	**8.5**
Malware Generation (MG)	3.32%	19.88%	15.52%	**50.66%**	30.2	22.1	24.3	**15.8**
Physical Harm (PH)	2.98%	24.31%	23.43%	**74.76%**	30.7	20.1	20.5	**7.3**
Economic Harm (EH)	5.68%	4.91%	8.91%	**72.04%**	24.02	24.16	23.43	**11.97**
Fraud (FR)	3.17%	18.56%	14.71%	**50.96%**	24.47	21.68	22.38	**15.80**
Pornography (PO)	4.14%	20.94%	19.11%	**69.84%**	24.30	21.25	21.58	**12.37**
Political Lobbying (PL)	67.67%	79.11%	76.46%	**98.38%**	18.71	14.43	16.06	**13.78**
Privacy Violence (PV)	8.97%	10.50%	12.97%	**81.79%**	27.03	24.94	21.98	**12.31**
Legal Opinion (LO)	74.56%	85.73%	86.52%	**100%**	16.97	8.25	**7.30**	7.74
Financial Advice (FA)	84.33%	88.12%	90.93%	**100%**	9.83	5.20	0.99	**5.77**
Health Consultation (HC)	76.50%	93.94%	91.22%	**96.00%**	16.04	8.41	10.04	**4.85**
Government Decision (GD)	90.29%	91.75%	91.25%	**98.72%**	13.73	11.88	11.39	**6.32**
Average	32.71%	43.38%	41.56%	**79.62%**	23.04	17.99	18.22	**10.03**

Highly Transferable Across Multiple MLLMs

Scenario	Case 1				Case 2			
	MiniGPT-4 (Base)	LLaVa	Qwen	GPT-4o	MiniGPT-4 (Base)	LLaVa	Qwen	Gemini
Illegal Activity (IA)	70.0%	66.5%	64.0%	45.5%	68.0%	64.5%	62.5%	47.5%
Hate Speech (HS)	66.0%	62.7%	60.7%	42.9%	64.0%	60.8%	58.9%	44.8%
Malware Generation (MG)	48.0%	45.6%	44.2%	31.2%	46.0%	43.7%	42.3%	32.2%
Physical Harm (PH)	72.0%	68.4%	66.2%	46.8%	70.0%	66.5%	64.4%	49.0%
Economic Harm (EH)	70.0%	66.5%	64.4%	45.5%	68.0%	64.5%	62.6%	47.5%
Fraud (FR)	48.0%	45.6%	44.2%	31.2%	46.0%	43.7%	42.3%	32.2%
Pornography (PO)	67.0%	63.7%	61.6%	43.6%	65.0%	61.8%	59.8%	45.5%
Political Lobbying (PL)	95.0%	90.3%	87.4%	61.8%	93.0%	88.4%	85.6%	65.1%
Privacy Violence (PV)	79.0%	75.1%	72.7%	51.4%	77.0%	73.2%	70.8%	53.9%
Legal Opinion (LO)	97.0%	92.2%	89.2%	63.1%	95.0%	90.3%	87.4%	66.5%
Financial Advice (FA)	97.0%	92.2%	89.2%	63.1%	95.0%	90.3%	87.4%	66.5%
Health Consultation (HC)	93.0%	88.4%	85.6%	60.5%	91.0%	86.5%	83.7%	63.7%
Government Decision (GD)	96.0%	91.2%	88.3%	62.4%	94.0%	89.3%	86.5%	65.8%
Average	75.5%	71.7%	69.1%	49.0%	73.6%	69.8%	67.3%	51.2%