Stealthy Porn: Understanding Real-World Adversarial Images for Illicit Online Promotion

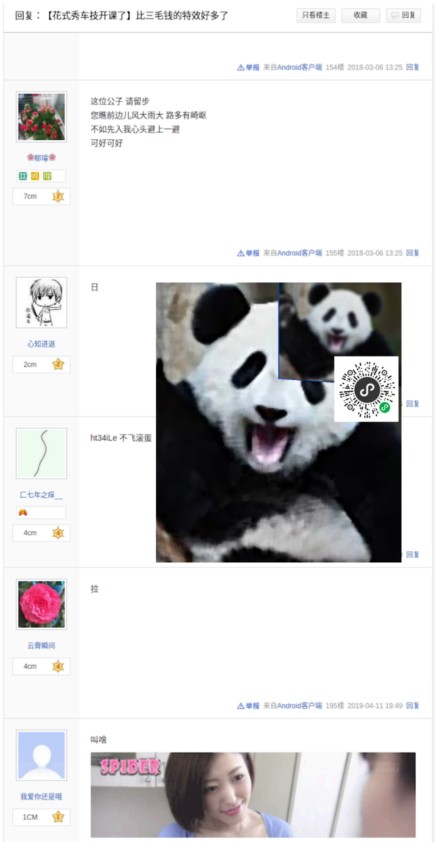
Kan Yuan, **Di Tang**, Xiaojing Liao, XiaoFeng Wang, Xuan Feng, Yi Chen, Menghan Sun, Haoran Lu, Kehuan Zhang

INDIANA UNIVERSITY

香港中文大學
The Chinese University of Hong Kong

中國科学院
CHINESE ACADEMY OF SCIENCES

A Case in the Wild



回复：【花式秀车技开课了】比三毛钱的特效好多了 只看楼主 收藏 回复

⚠举报 来自Android客户端 154楼 2018-03-06 13:25 回复

❀郁琣❀ 7cm

这位公子 请留步
您瞧前边儿风大雨大 路多有崎岖
不如先入我心头避上一避
可好可好

⚠举报 来自Android客户端 155楼 2018-03-06 13:25 回复

心知进退 2cm

日

匸七年之痒__ 4cm

ht34iLe 不飞滚蛋

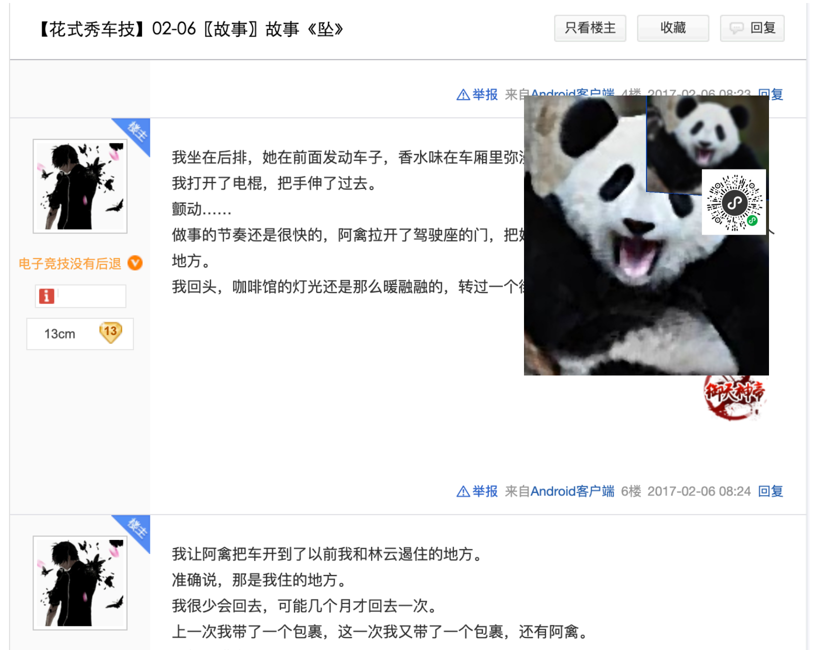
云雩瞬间 4cm

拉

⚠举报 来自Android客户端 195楼 2019-04-11 19:49 回复

我爱你还是哦 1CM

叫啥



【花式秀车技】02-06〖故事〗故事《坠》 只看楼主 收藏 回复

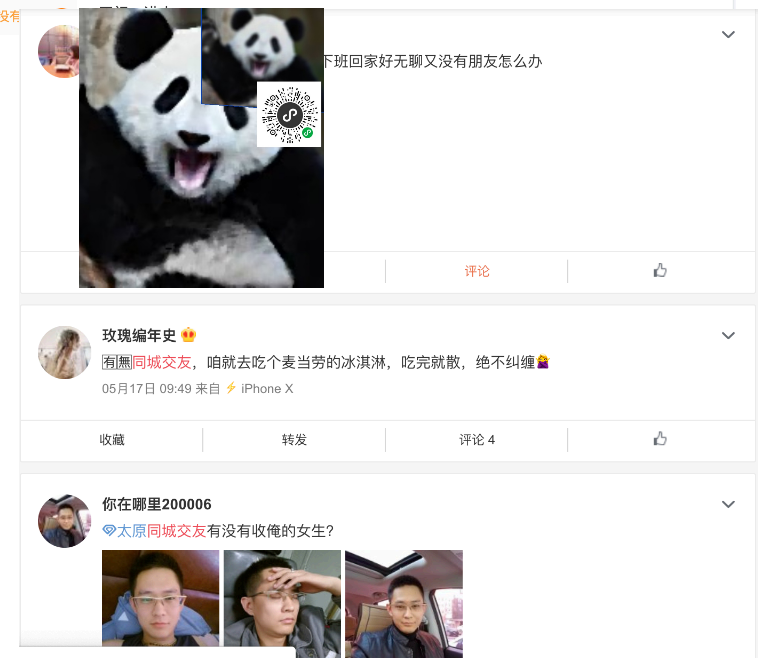
⚠举报 来自Android客户端 4楼 2017-02-06 08:23 回复

电子竞技没有后退 13cm

我坐在后排，她在前面发动车子，香水味在车厢里弥
我打开了电棍，把手伸了过去。
颤动……
做事的节奏还是很快的，阿禽拉开了驾驶座的门，把
地方。
我回头，咖啡馆的灯光还是那么暖融融的，转过一个

⚠举报 来自Android客户端 6楼 2017-02-06 08:24 回复

我让阿禽把车开到了以前我和林云遏住的地方。
准确说，那是我住的地方。
我很少会回去，可能几个月才回去一次。
上一次我带了一个包裹，这一次我又带了一个包裹，还有阿禽。



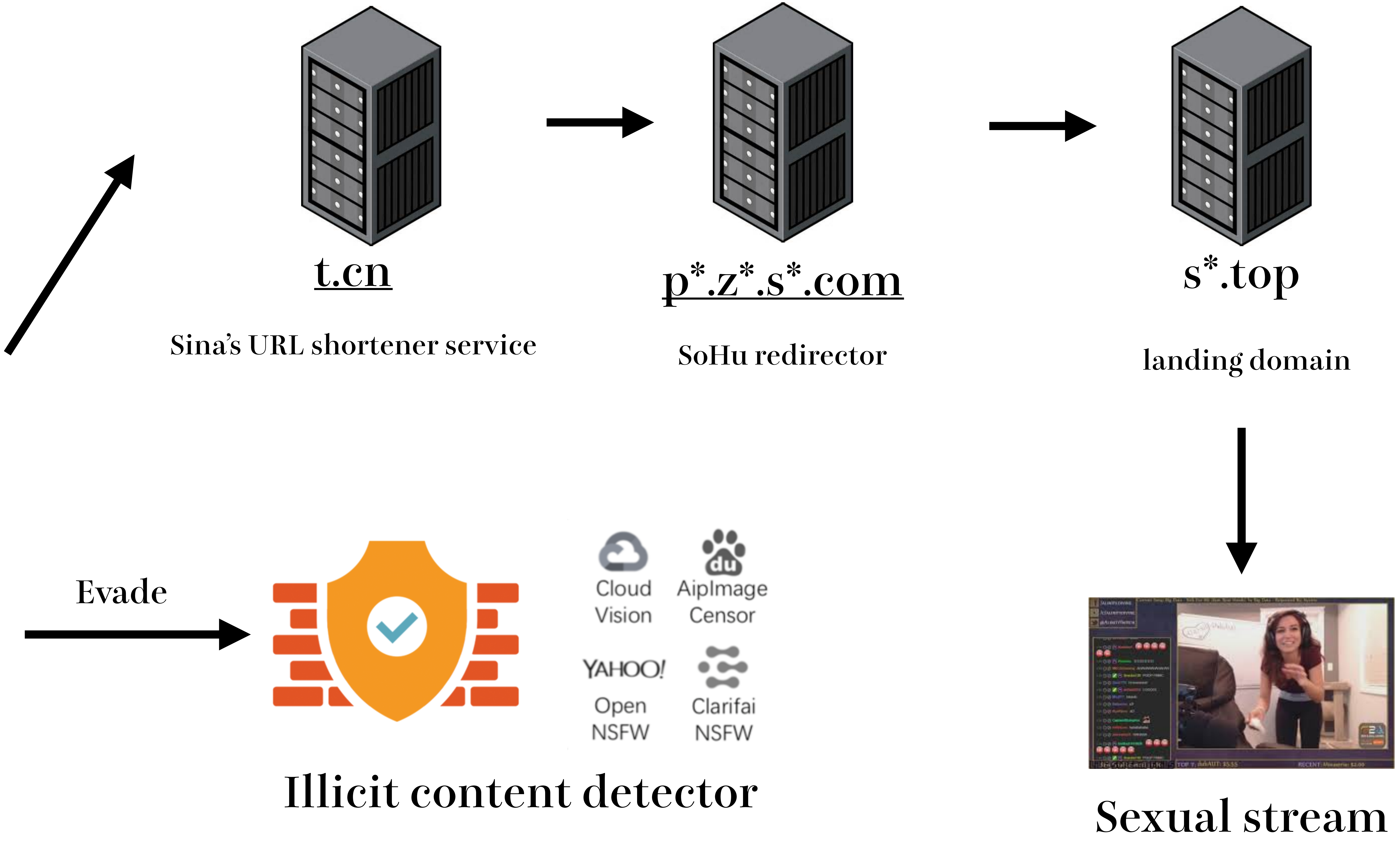
下班回家好无聊又没有朋友怎么办

评论

玫瑰编年史
有無同城交友，咱就去吃个麦当劳的冰淇淋，吃完就散，绝不纠缠
05月17日 09:49 来自 ⚡ iPhone X

收藏 转发 评论 4

你在哪里200006
太原同城交友有没有收俺的女生?



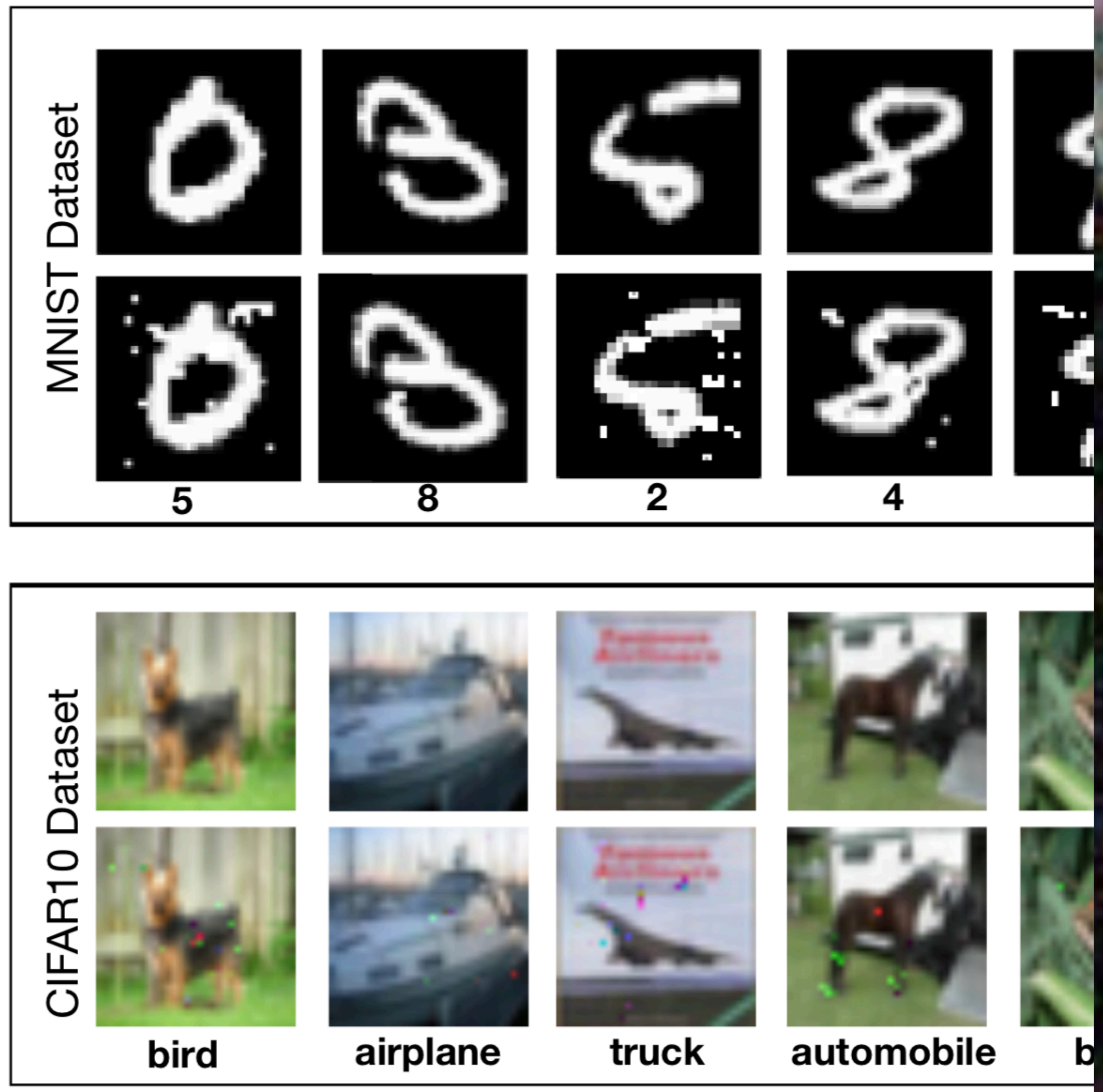
Adversarial Images in the Real-World

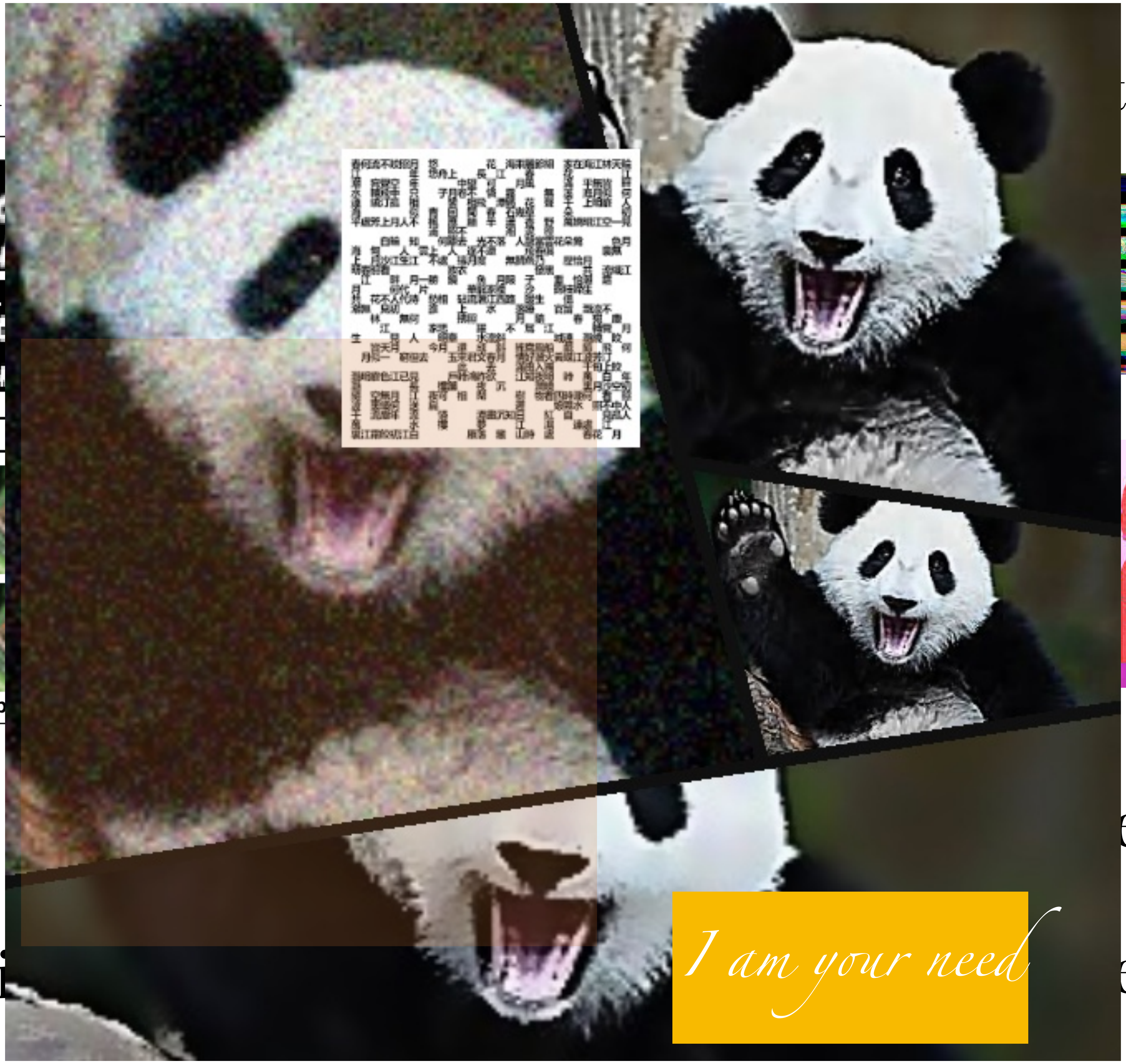
APPIs

Adversarial Promotional Porn Images

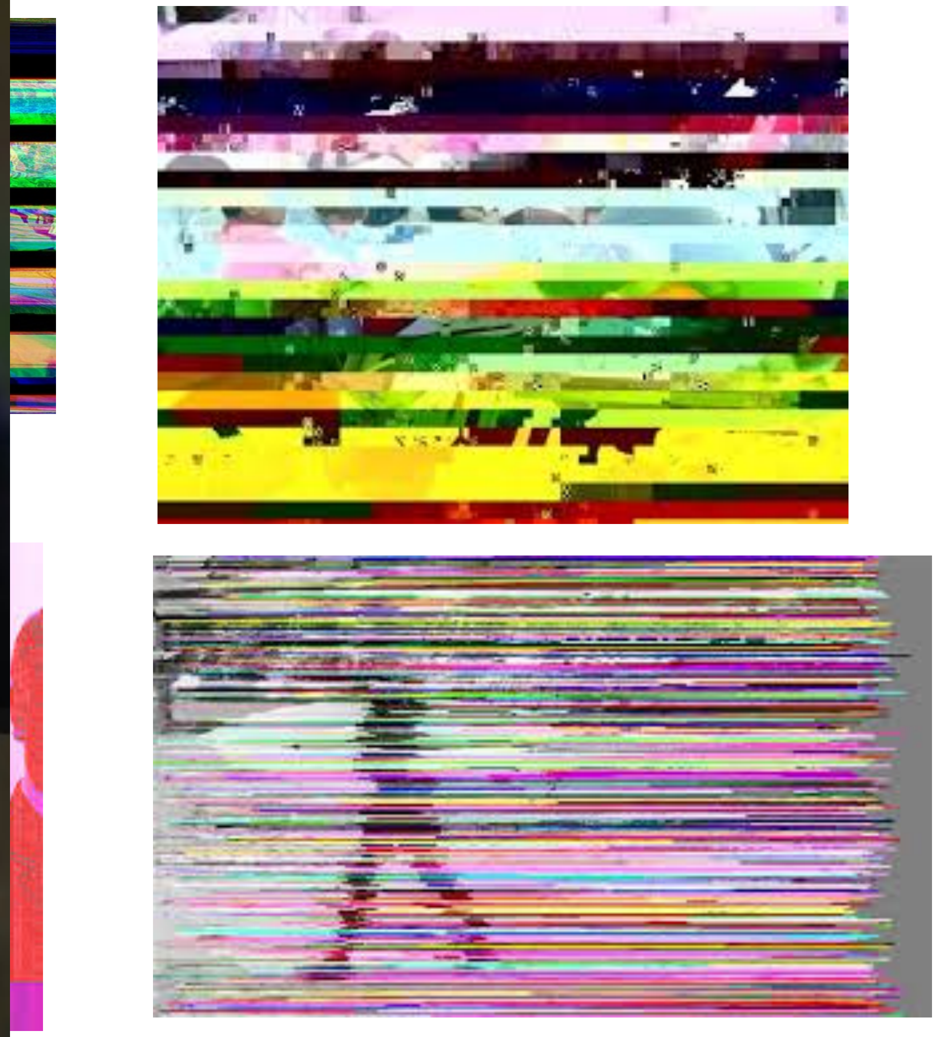
APPIs are Different

Adversarial Exampl





ted Images

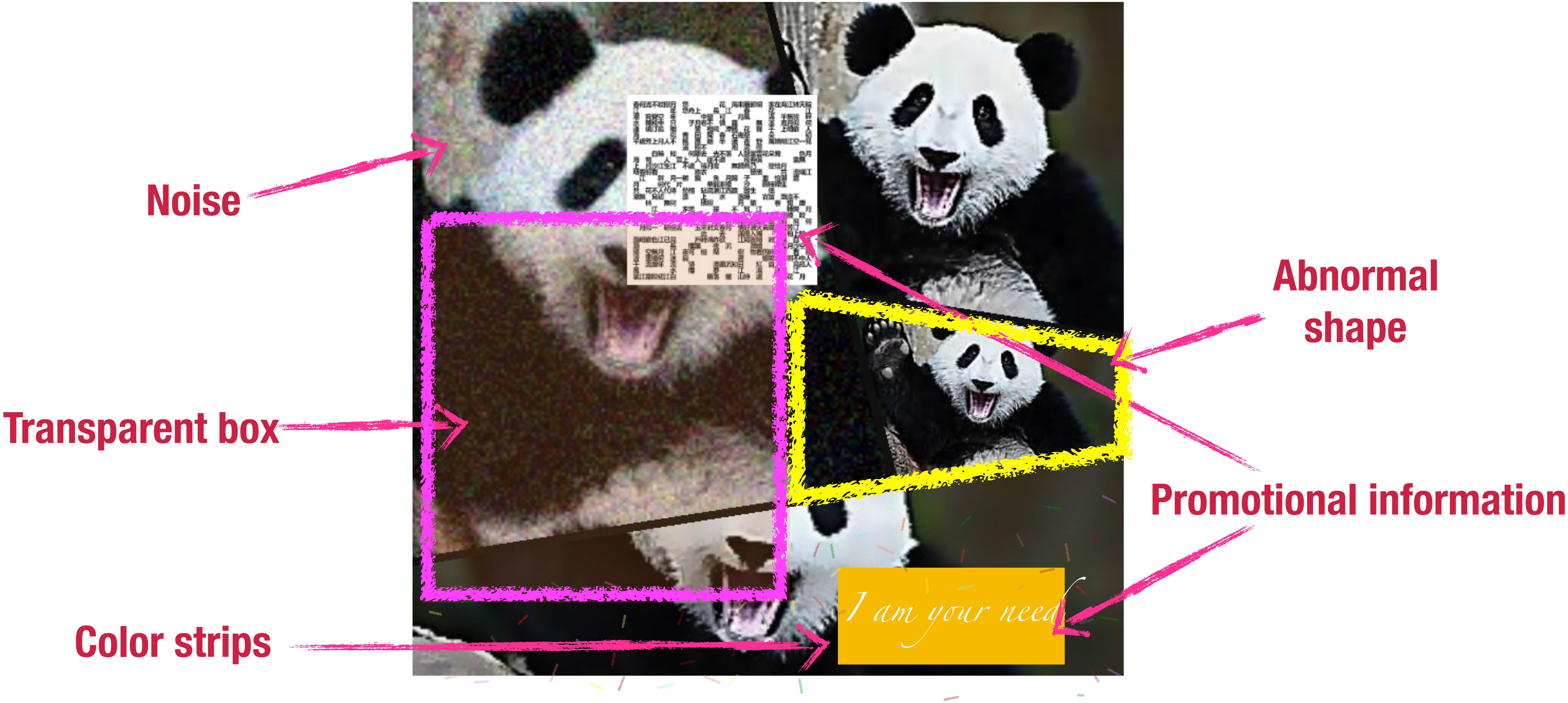


- small noise
- nearly indistingui

e

ess

APPIs Composition



Noise

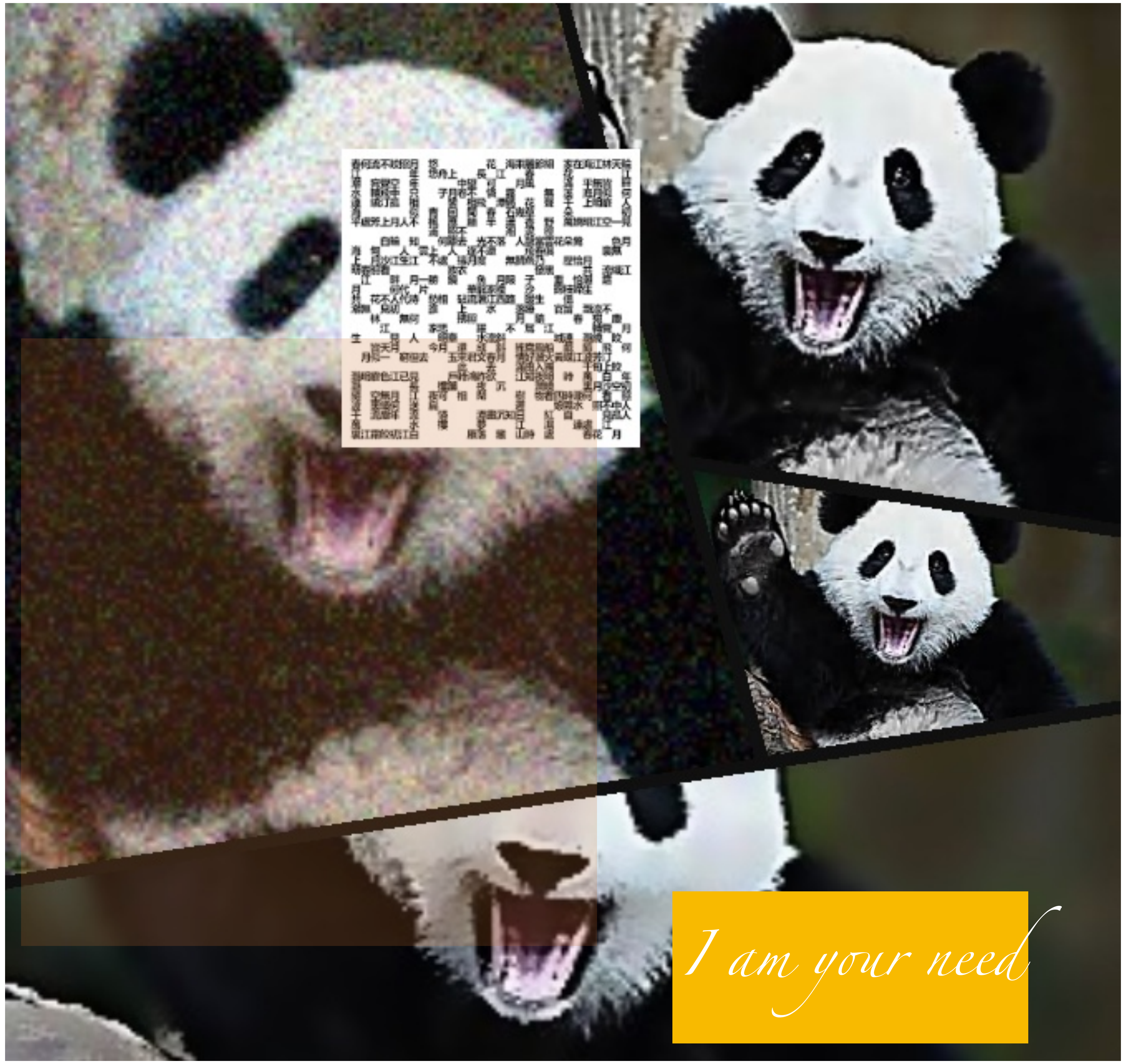
Transparent box

Color strips

Abnormal shape

Promotional information

Malèna: Finding Stealthy Porn



Two common characters:

- Promotional content.
- less obfuscated explicit content

Malèna: Finding Stealthy Porn

promotional content

evasiveness

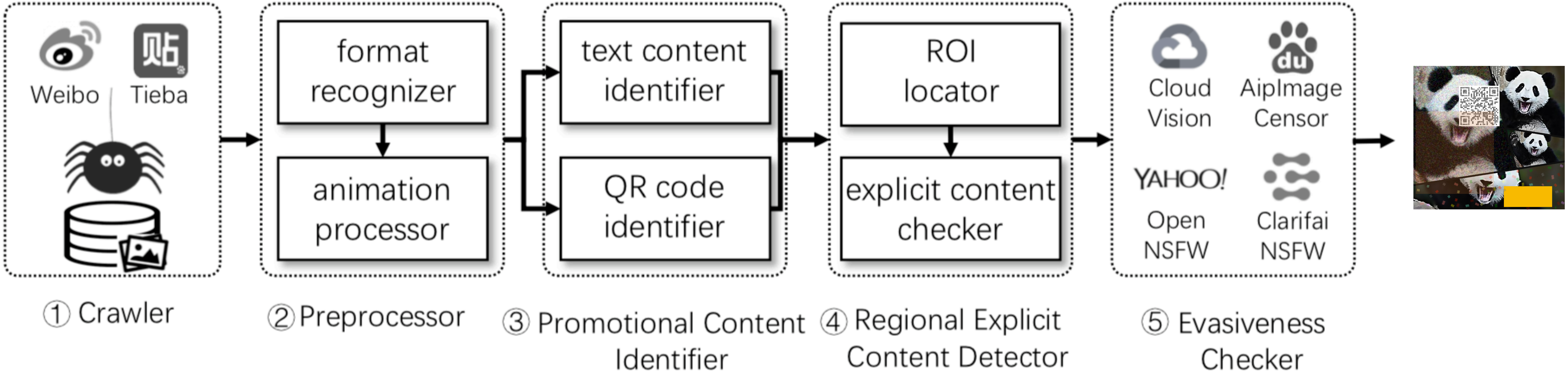


image type

explicit content

APPIs

Malèna: Finding Stealthy Porn

promotional content

evasiveness

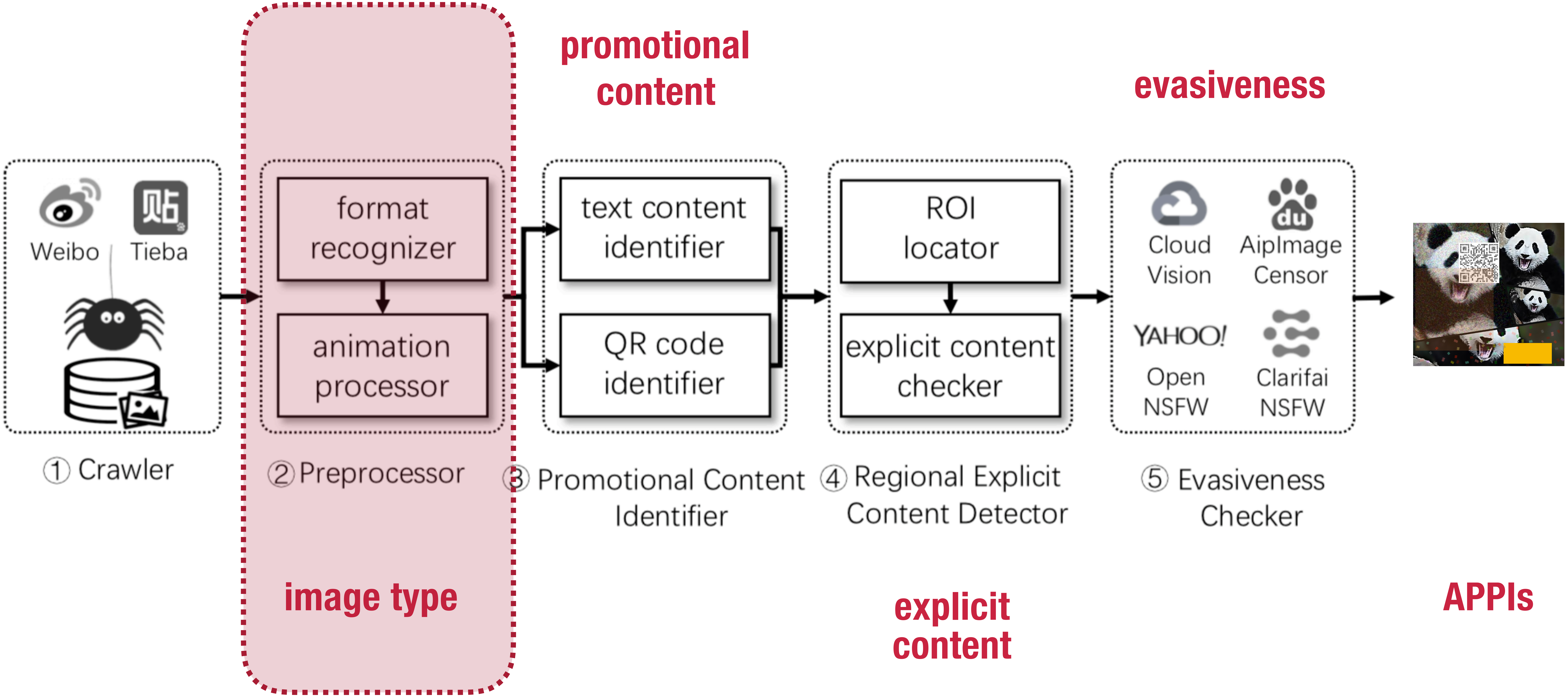
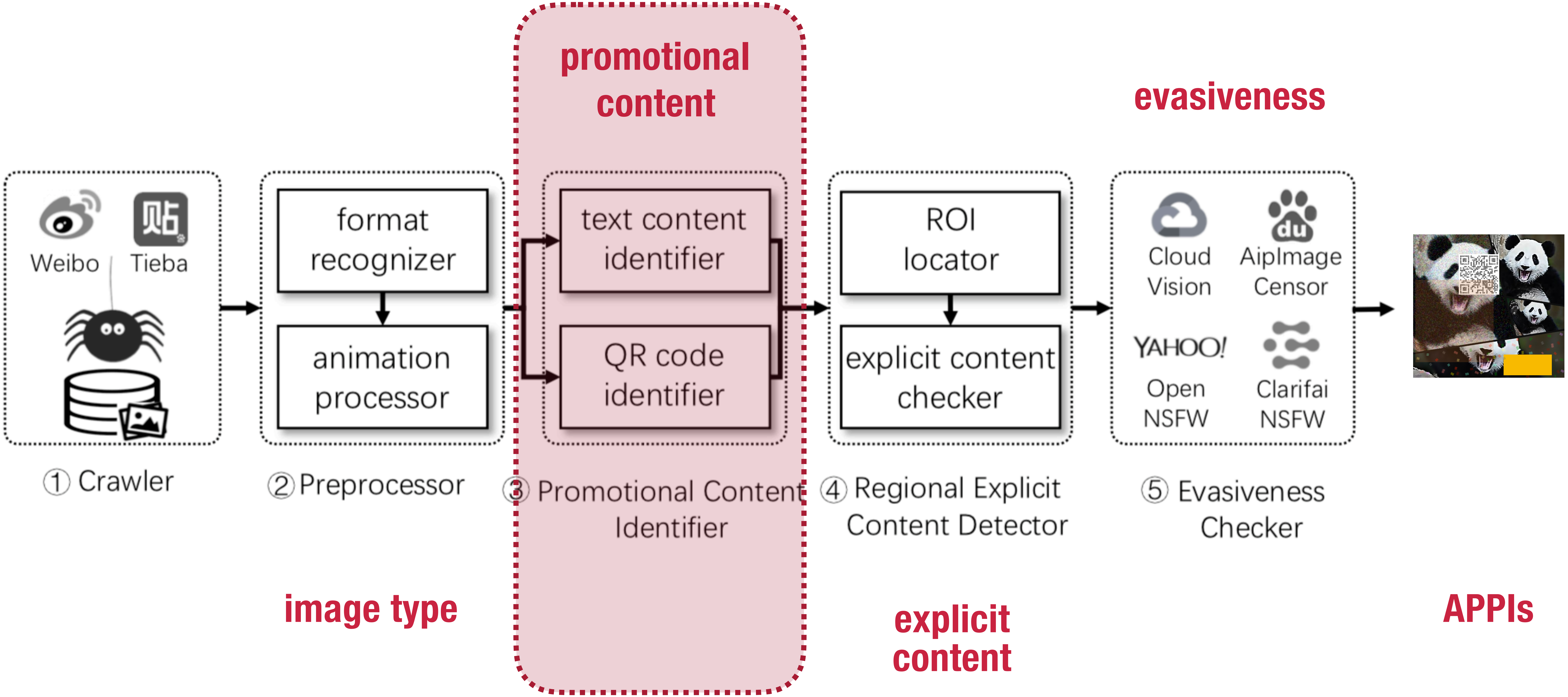


image type

explicit content

APPIs

Malèna: Finding Stealthy Porn



promotional content

evasiveness

Weibo

Tieba

format recognizer

animation processor

text content identifier

QR code identifier

ROI locator

explicit content checker

Cloud Vision

AipImage Censor

YAHOO! Open NSFW

Clarifai NSFW

① Crawler

② Preprocessor

③ Promotional Content Identifier

④ Regional Explicit Content Detector

⑤ Evasiveness Checker

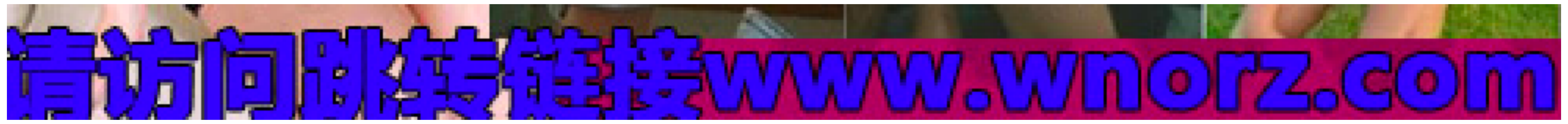
image type

explicit content

APPIs

Malèna: Promotional Content Identifier

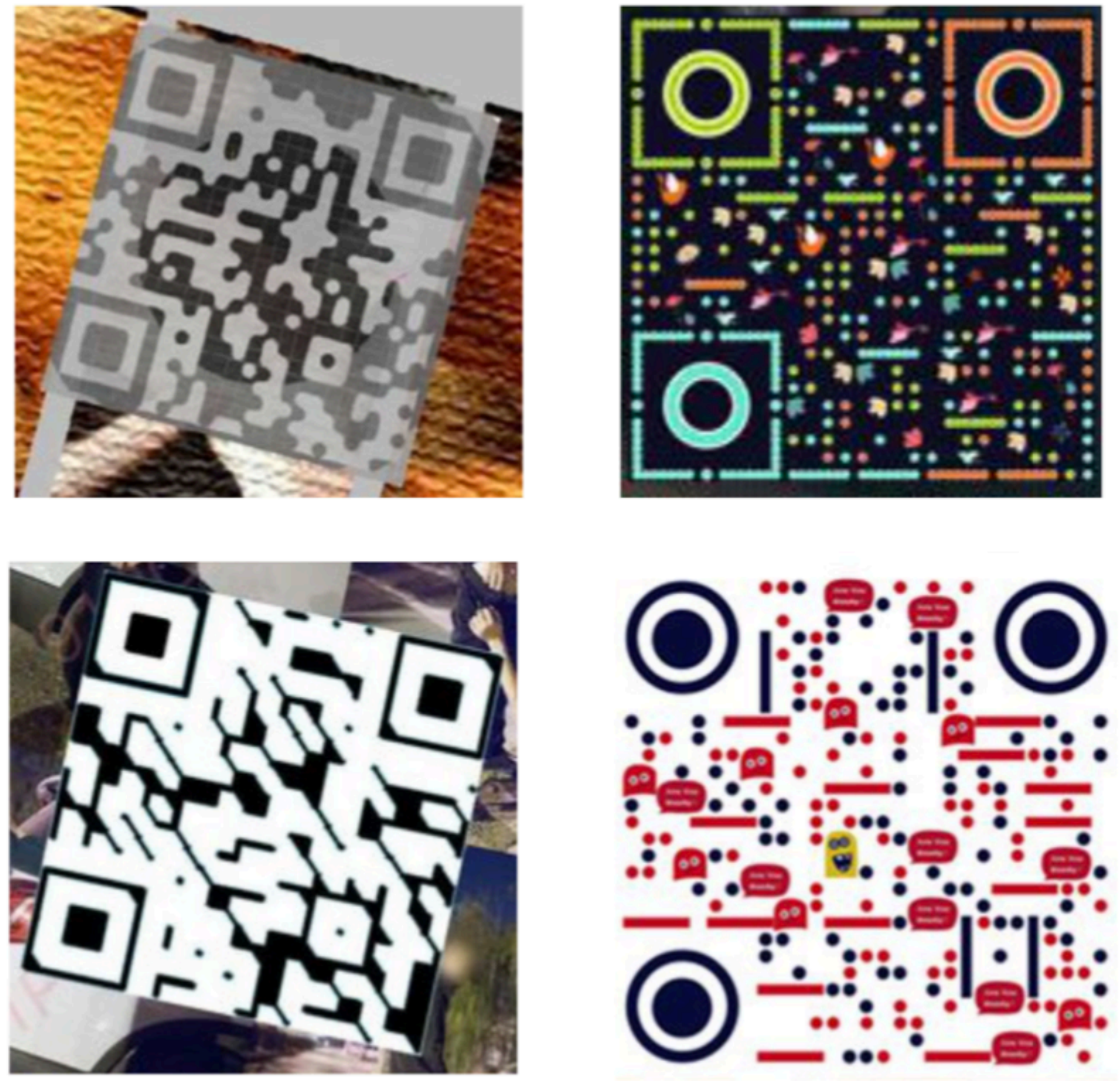
Text:



请访问跳转链接www.wnorz.com

✓ PiexlLink

QRcode:

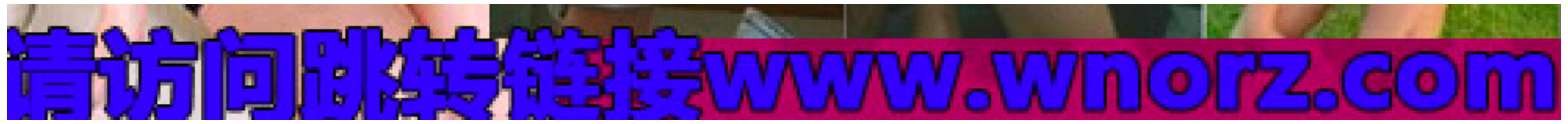


✗ ZBar, ZXing, BoofCV

✓ ~~WeChat~~

Malèna: Promotional Content Identifier

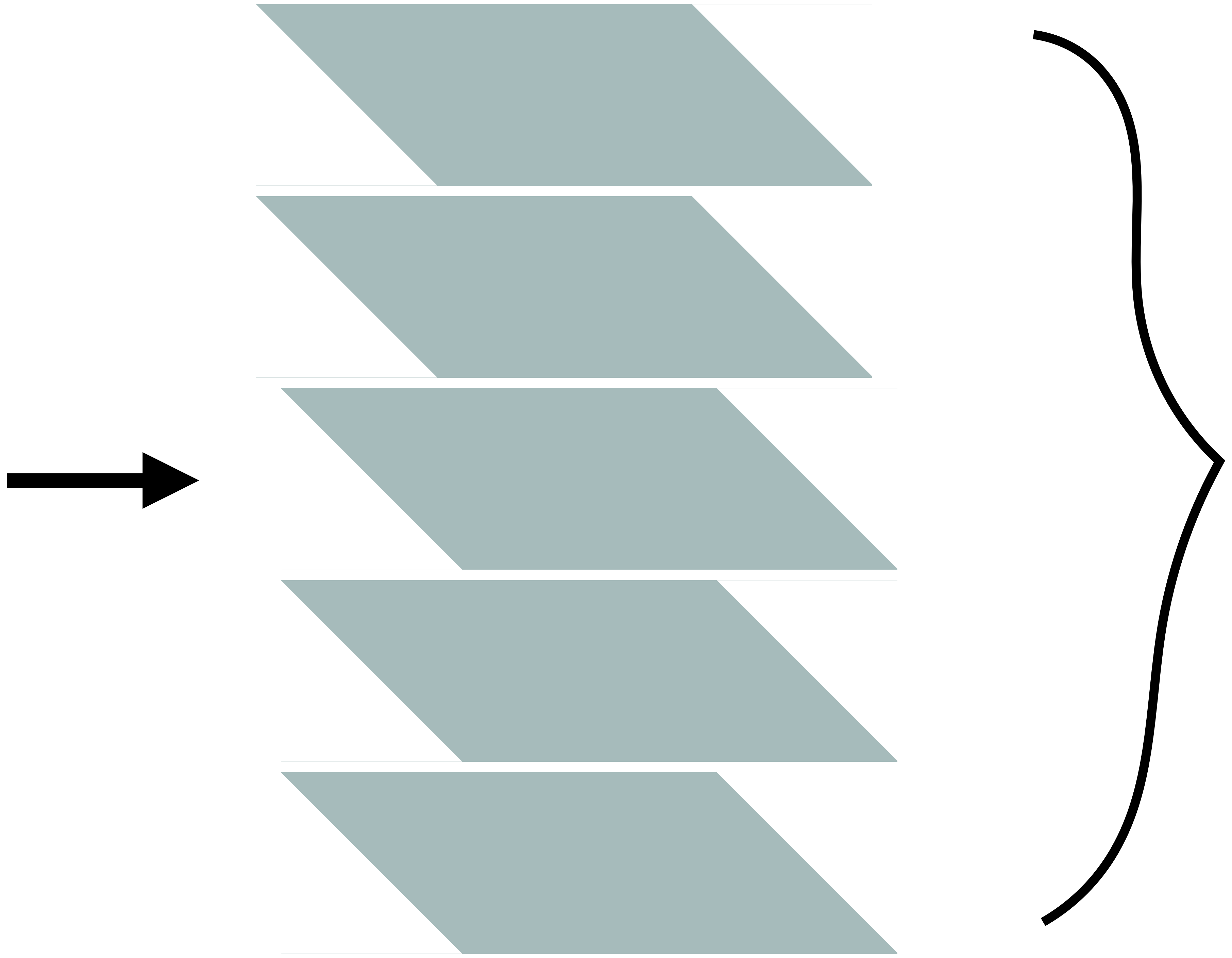
Text:



请访问跳转链接www.wnorz.com

✓ PiexlLink

QRcode:



✓

Malèna: Finding Stealthy Porn

promotional content

evasiveness

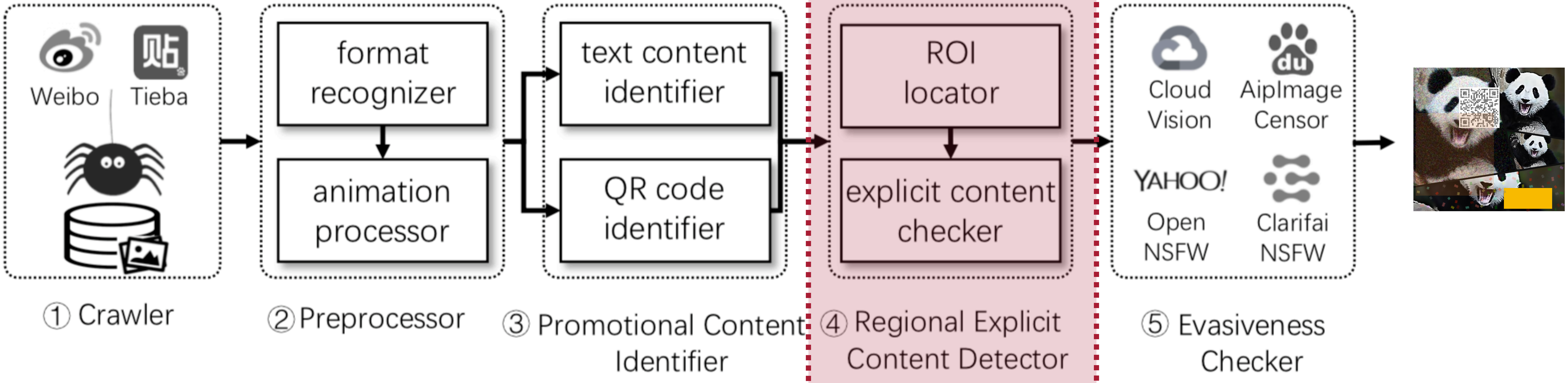


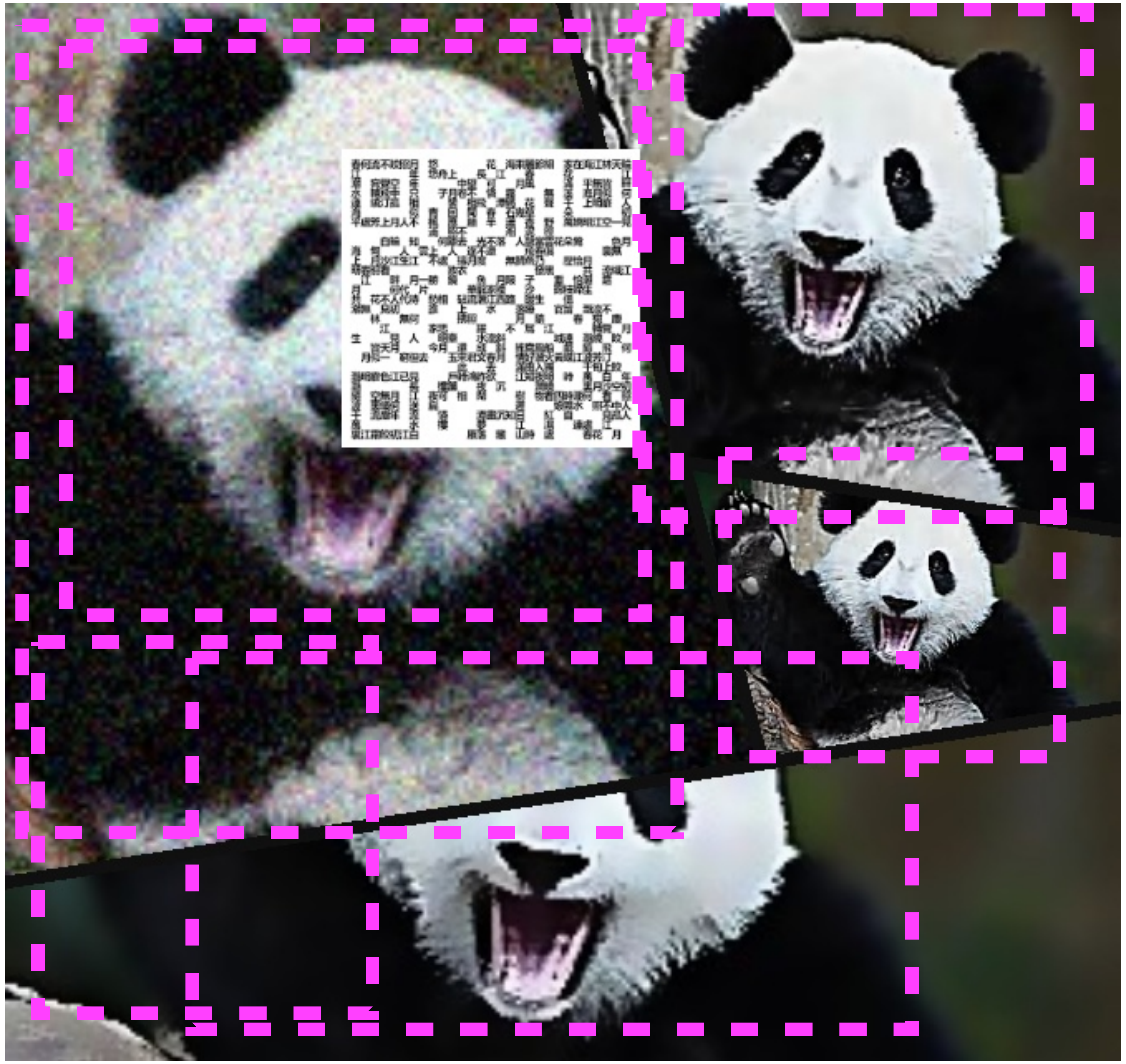
image type

explicit content

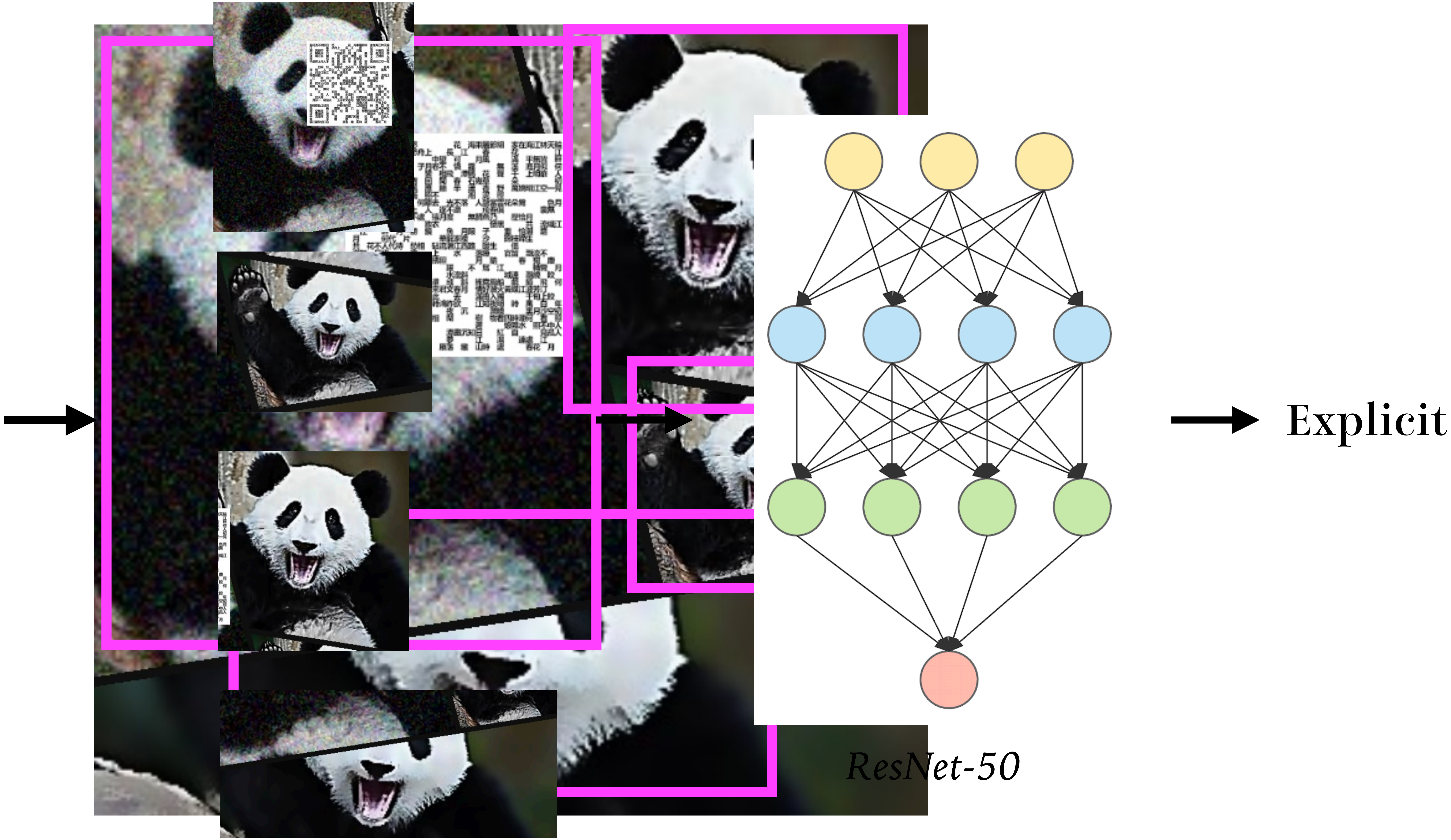
APPIs

Malèna: Regional Explicit Content Detector

Malèna: Regional Explicit Content Detector



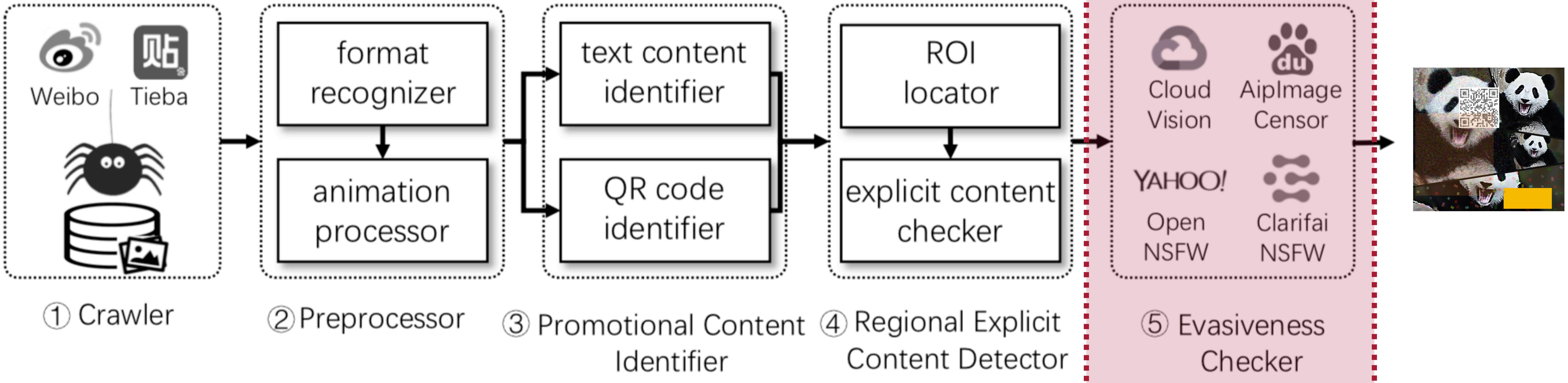
Malèna: Regional Explicit Content Detector



Malèna: Finding Stealthy Porn

promotional content

evasiveness



Weibo | Tieba

① Crawler

format recognizer

animation processor

② Preprocessor

text content identifier

QR code identifier

③ Promotional Content Identifier

ROI locator

explicit content checker

④ Regional Explicit Content Detector

Cloud Vision

AipImage Censor

YAHOO! Open NSFW

Clarifai NSFW

⑤ Evasiveness Checker

image type

explicit content

APPIs

MALÈNA: PERFORMANCE

- Performance: 91% precision , 85% recall

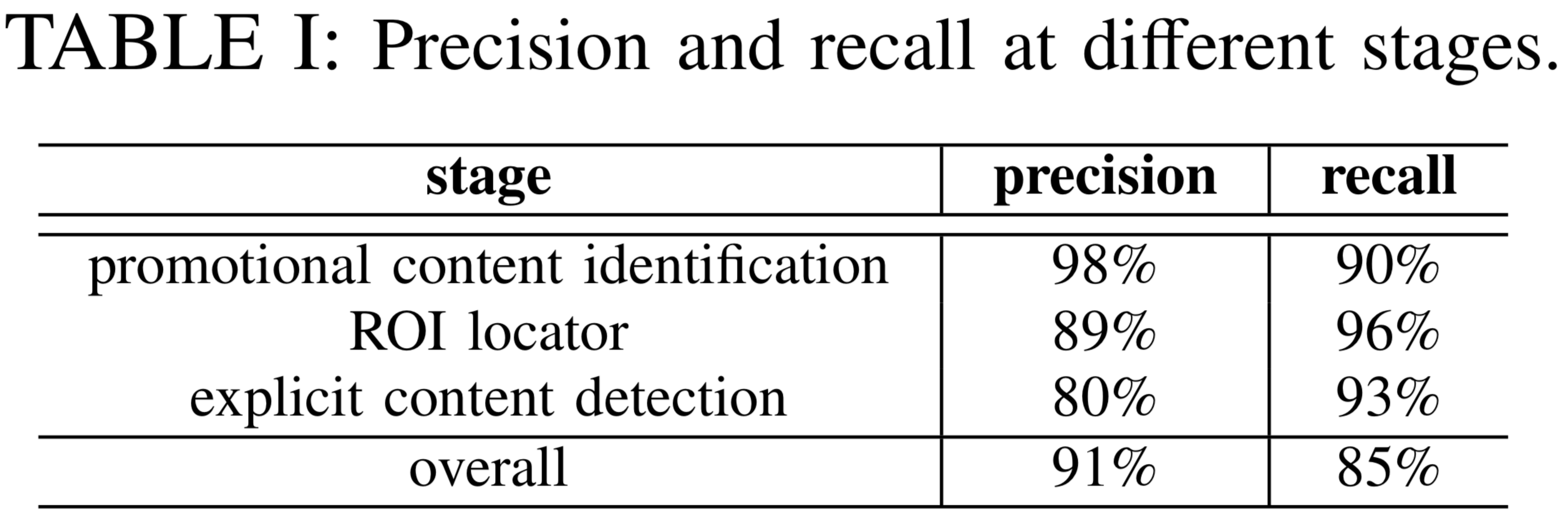


TABLE I: Precision and recall at different stages.

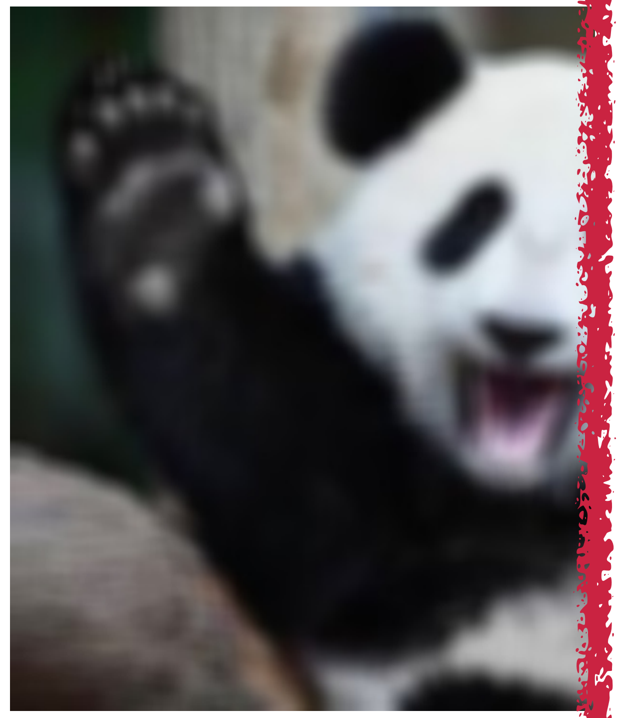
stage	**precision**	**recall**
promotional content identification	98%	90%
ROI locator	89%	96%
explicit content detection	80%	93%
overall	91%	85%

- Result: 4,353/6,163 APPIs , from 4M images, 76K posts (Baidu Tieba, Weibo)

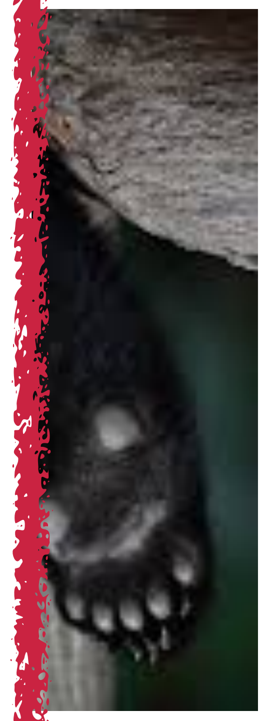
Measurement

- Visual pattern.
- Promotional content.
- Distribution channels.

Measurement: Visual Pattern



Blu



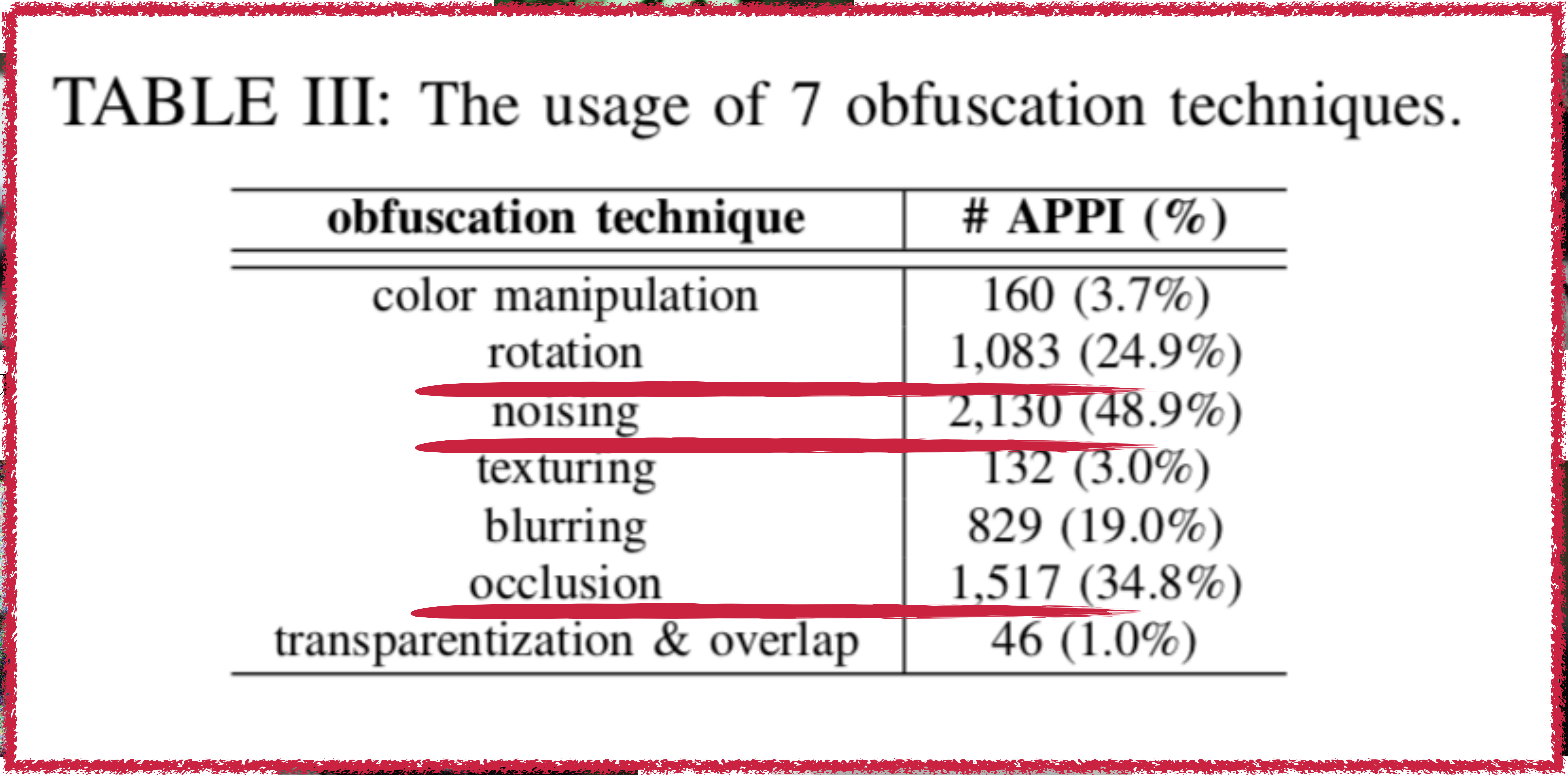
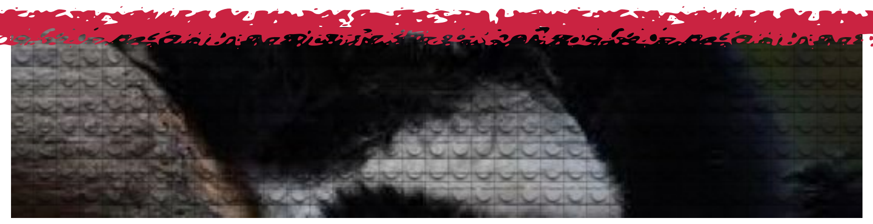


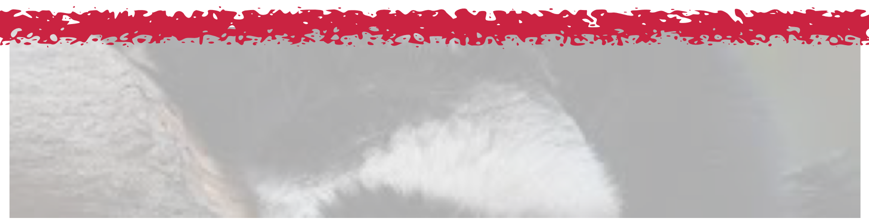
TABLE III: The usage of 7 obfuscation techniques.

obfuscation technique	# APPI (%)
color manipulation	160 (3.7%)
rotation	1,083 (24.9%)
noising	2,130 (48.9%)
texturing	132 (3.0%)
blurring	829 (19.0%)
occlusion	1,517 (34.8%)
transparentization & overlap	46 (1.0%)

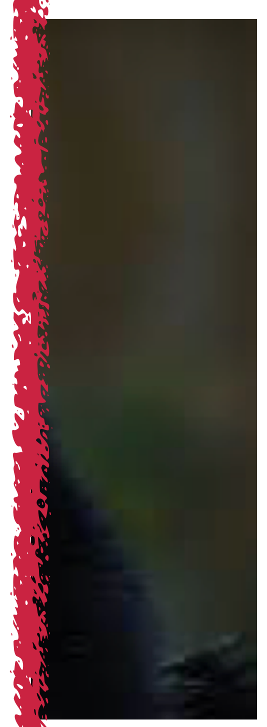
Noise



Texture



Transparentization



Occlusion

MEASUREMENT: VISUAL PATTERN

- Rotation
  - ➤ 45 and 135 degrees are effective

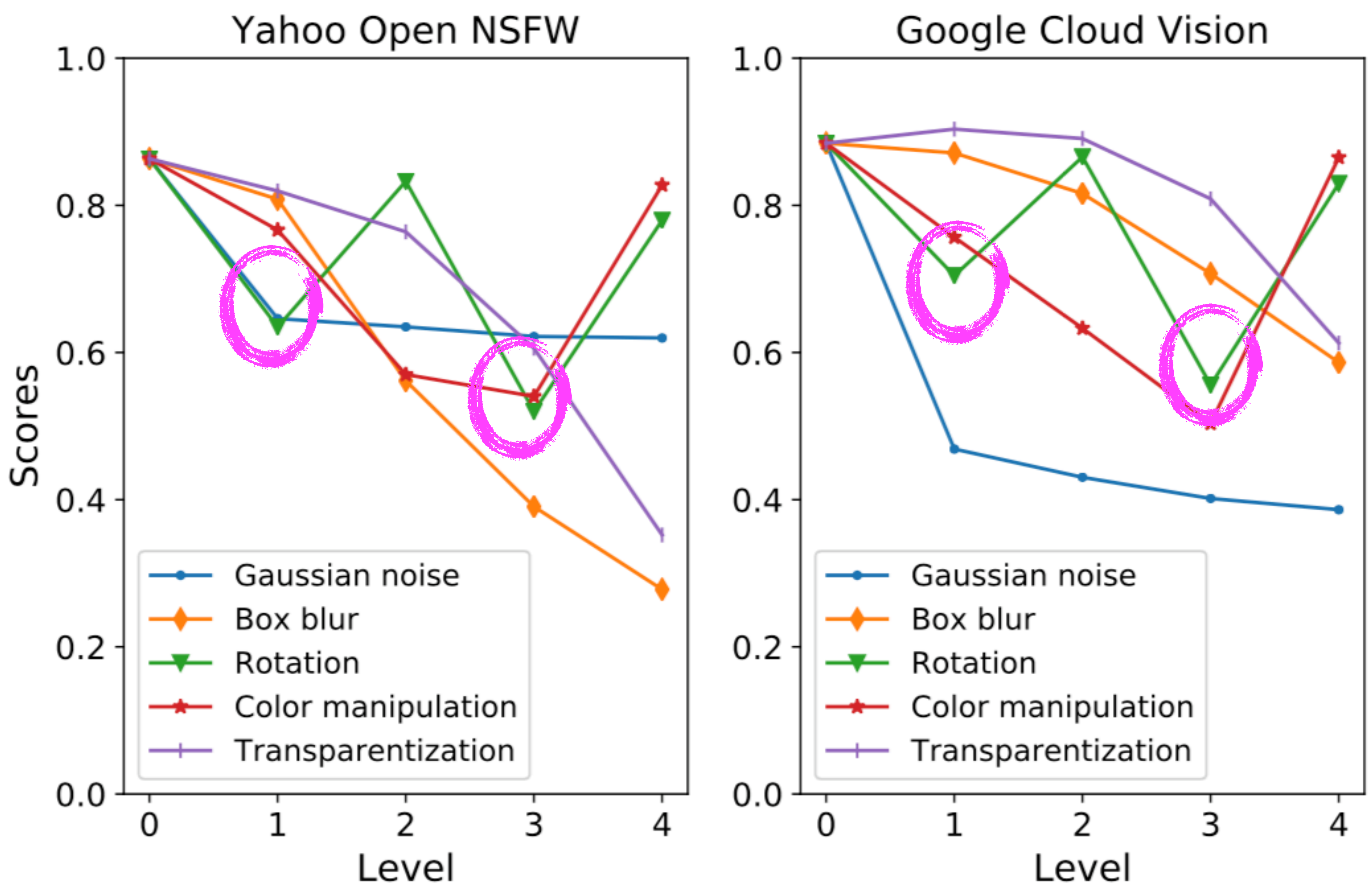


Fig. 10: Explicit content detection results on the distorted images.

Measurement: Visual Pattern

- Rotation
  - 45 and 135 degrees are effective
- Noising
  - Less noising is enough

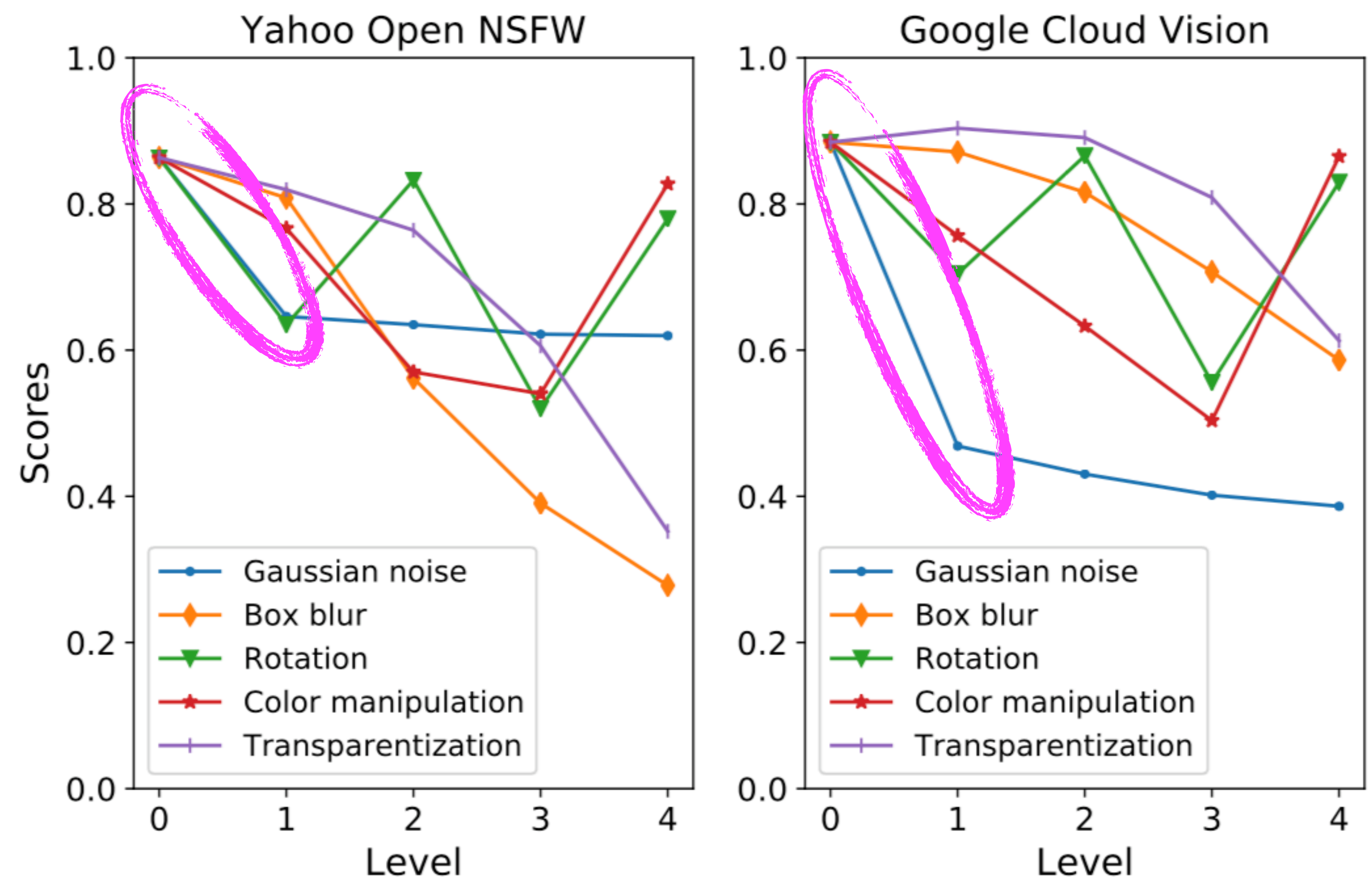


Fig. 10: Explicit content detection results on the distorted images.

Measurement: Visual Pattern

- Rotation
  - ➤ 45 and 135 degrees are effective
- Noising
  - ➤ Less noising is enough
- Color-manipulation
  - ➤ Green is evasive colour

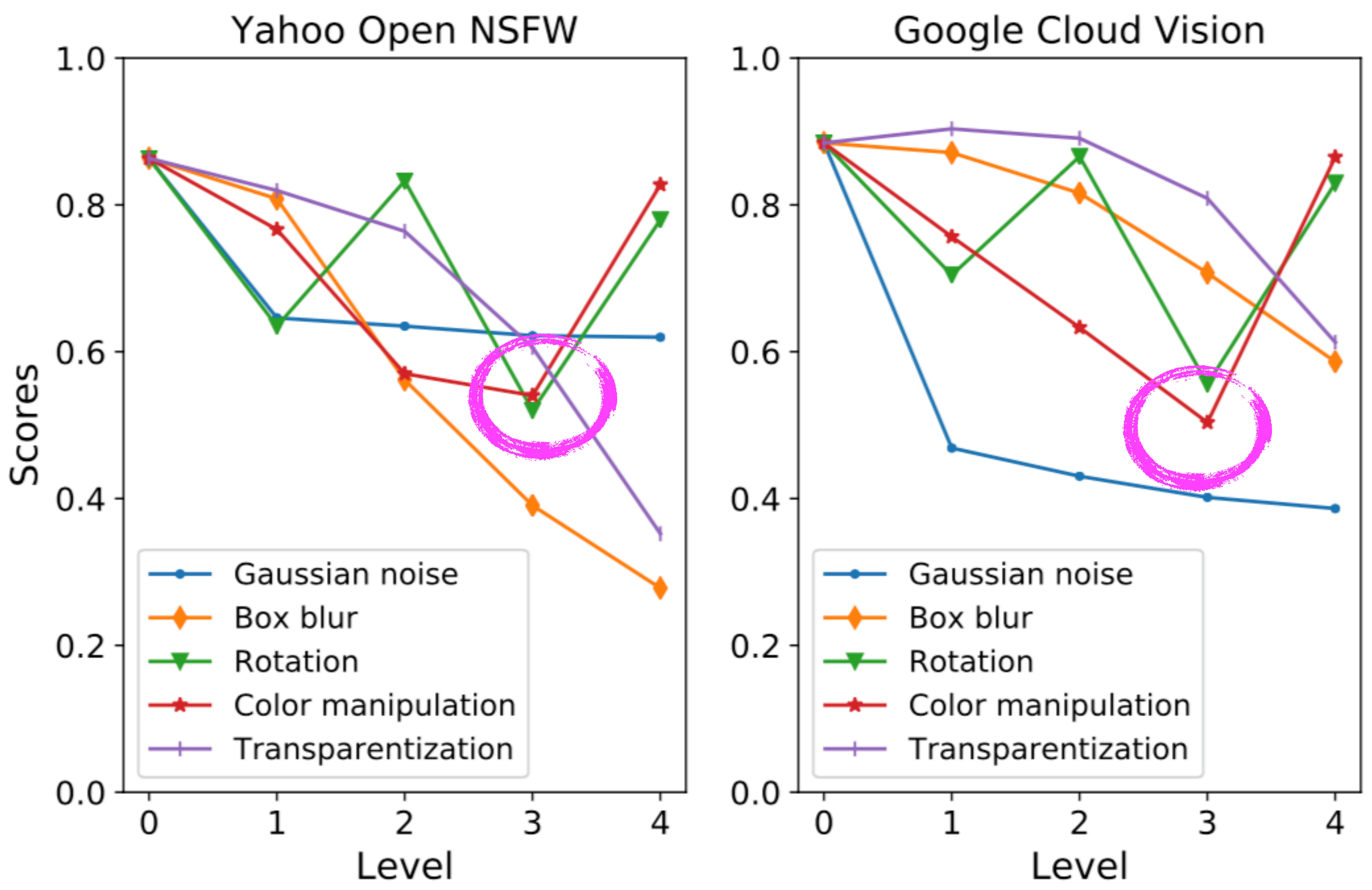
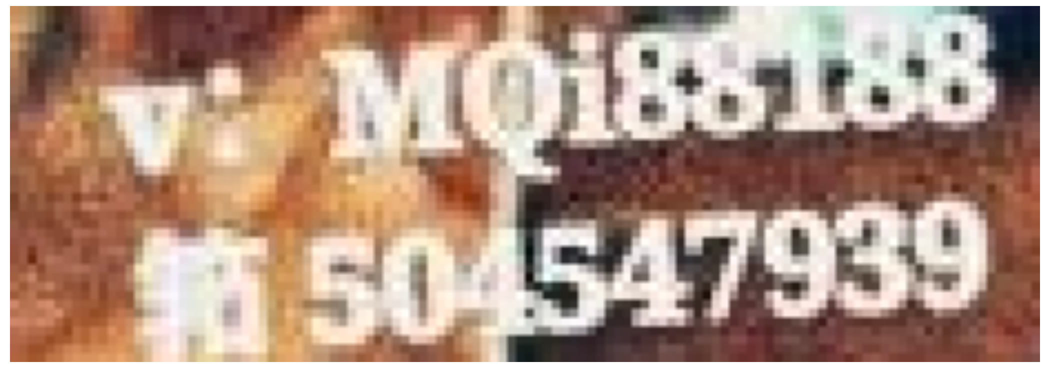
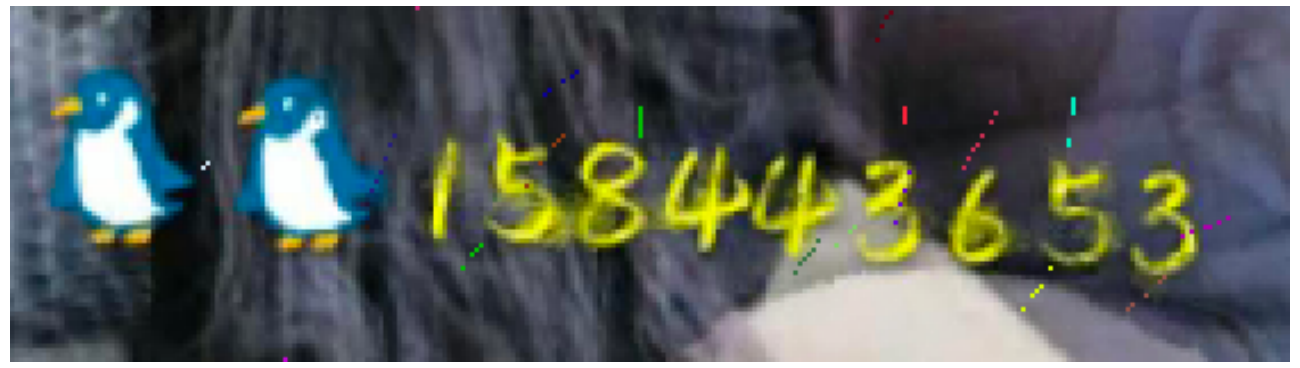


Fig. 10: Explicit content detection results on the distorted images.

Measurement: Promotional Content





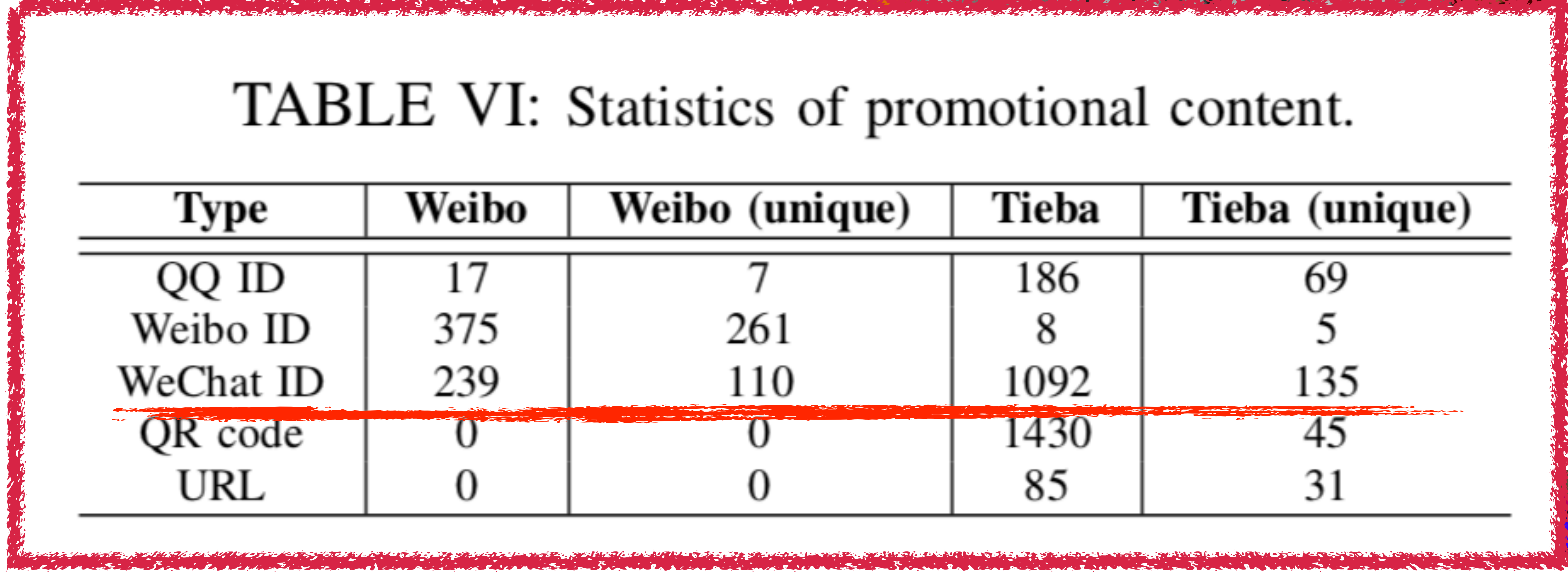
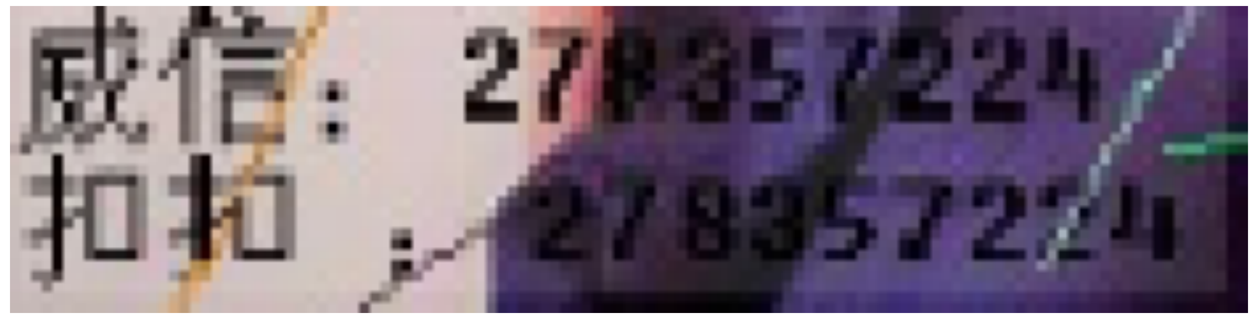
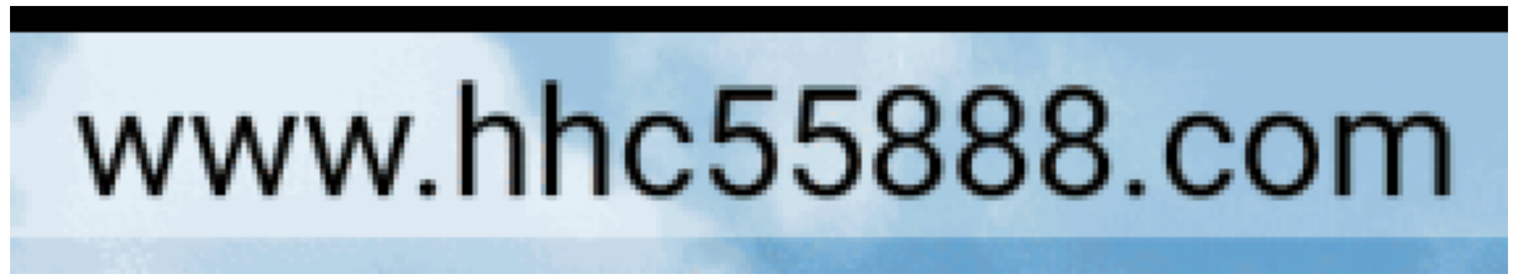
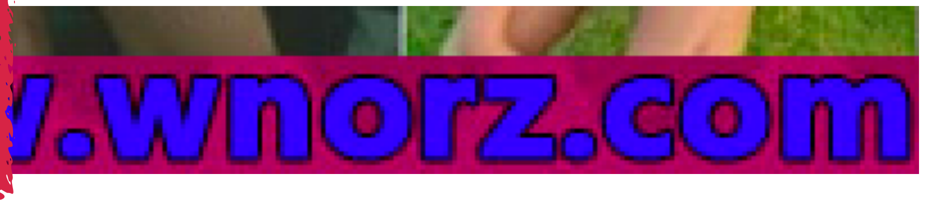


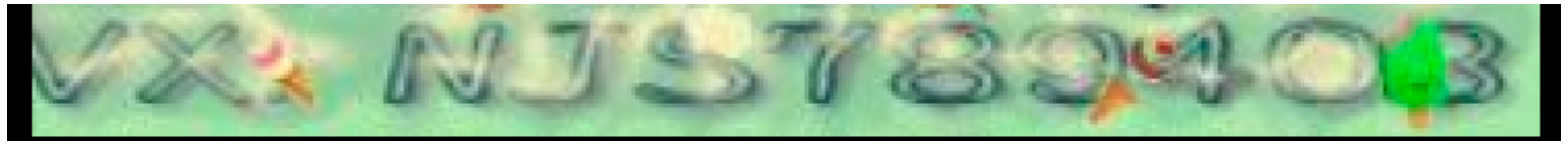
TABLE VI: Statistics of promotional content.

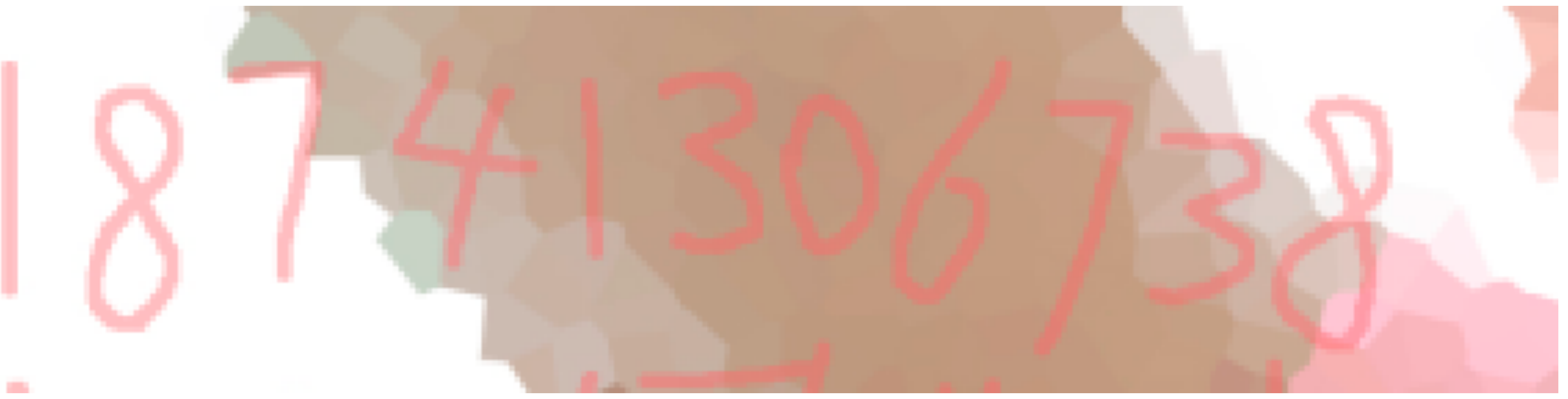
Type	Weibo	Weibo (unique)	Tieba	Tieba (unique)
QQ ID	17	7	186	69
Weibo ID	375	261	8	5
WeChat ID	239	110	1092	135
QR code	0	0	1430	45
URL	0	0	85	31



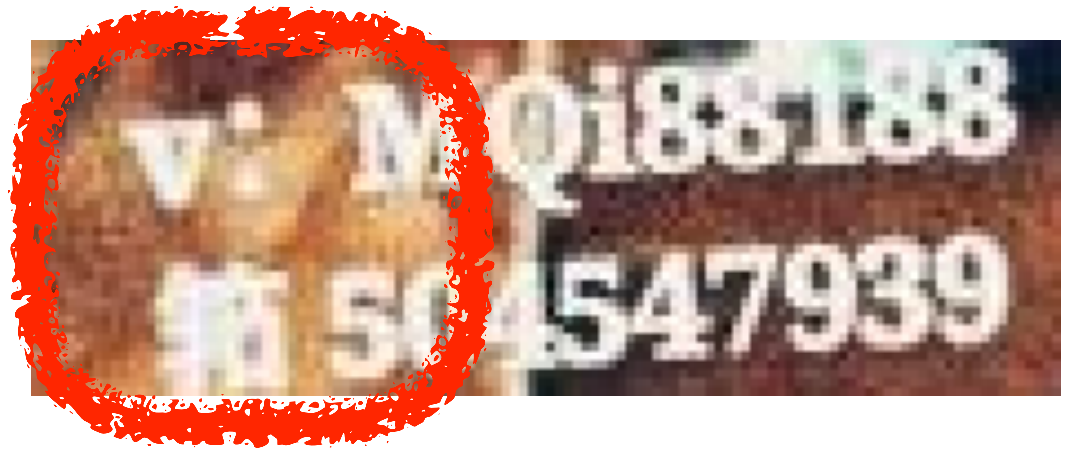


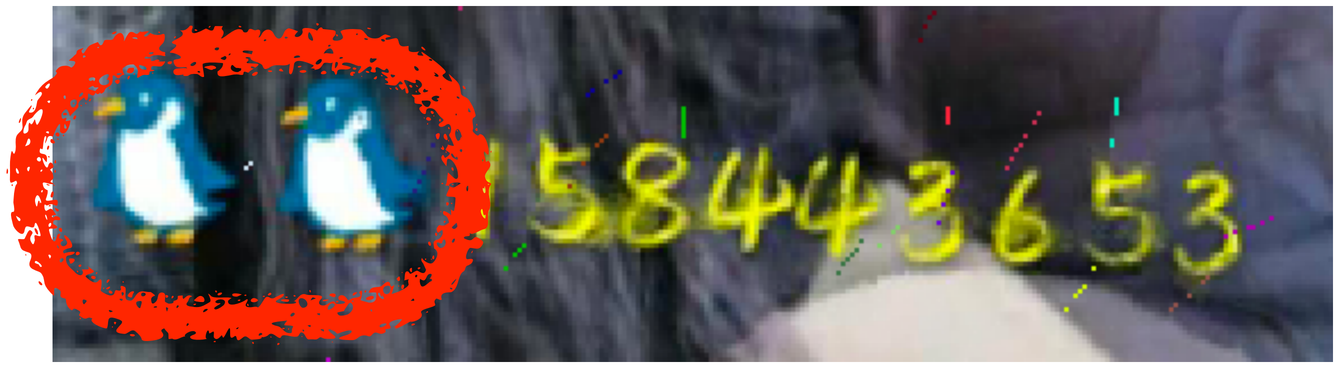






Measurement: Promotional Content





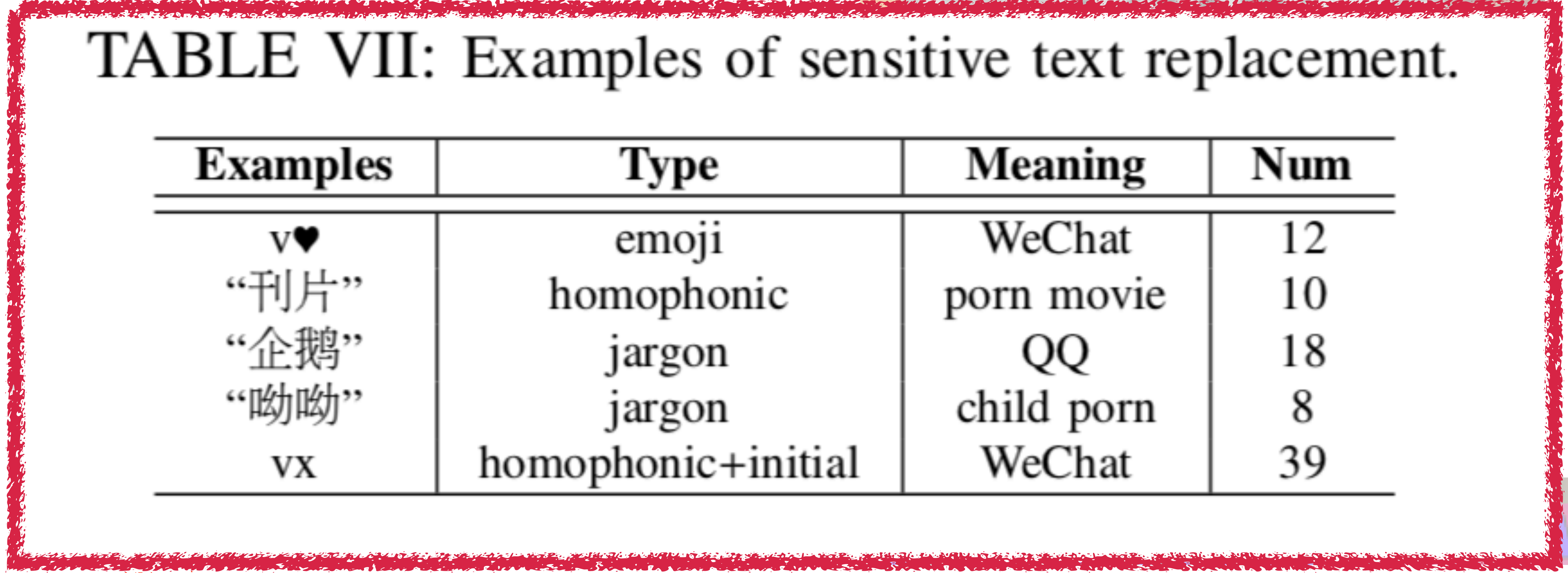
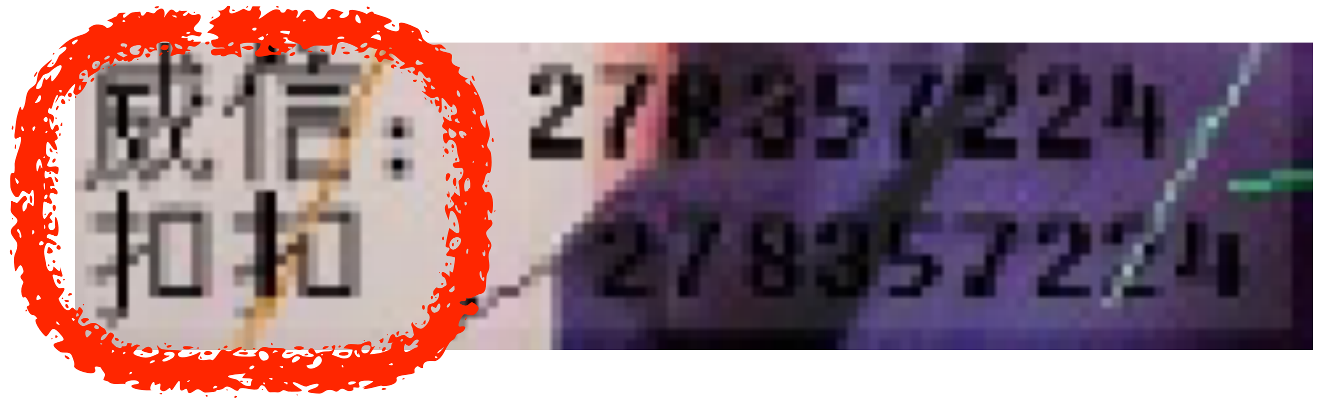
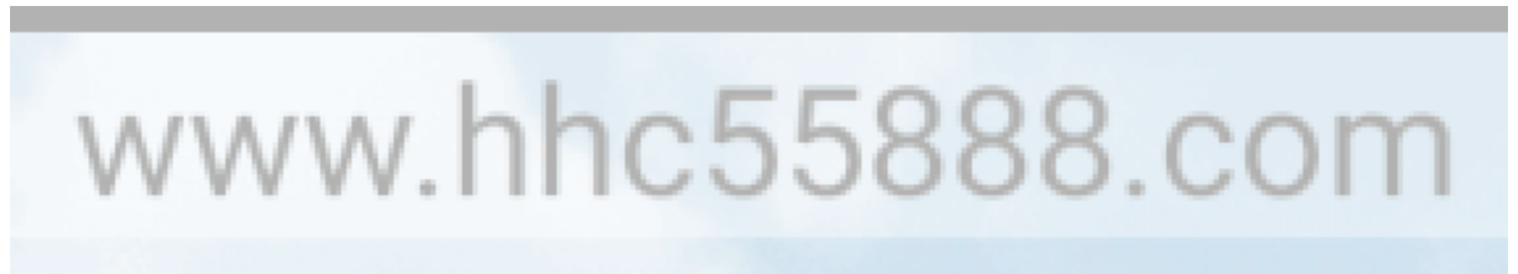
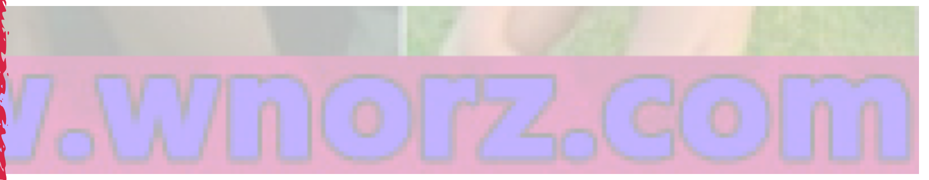


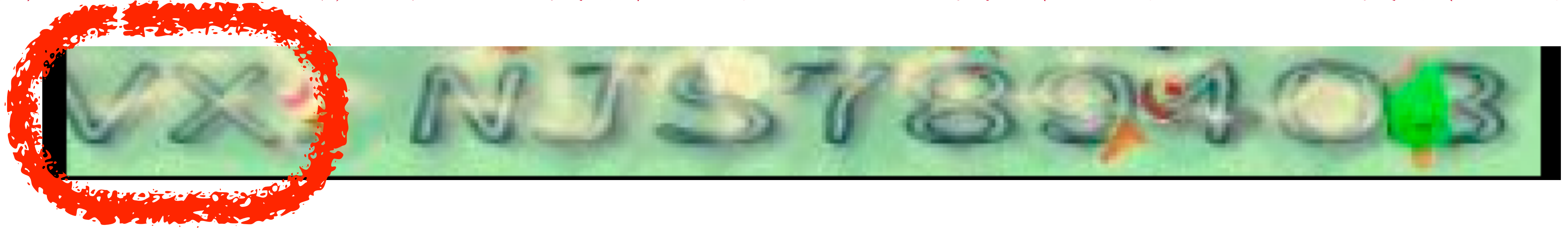
TABLE VII: Examples of sensitive text replacement.

Examples	Type	Meaning	Num
v♥	emoji	WeChat	12
“刊片”	homophonic	porn movie	10
“企鹅”	jargon	QQ	18
“呦呦”	jargon	child porn	8
vx	homophonic+initial	WeChat	39

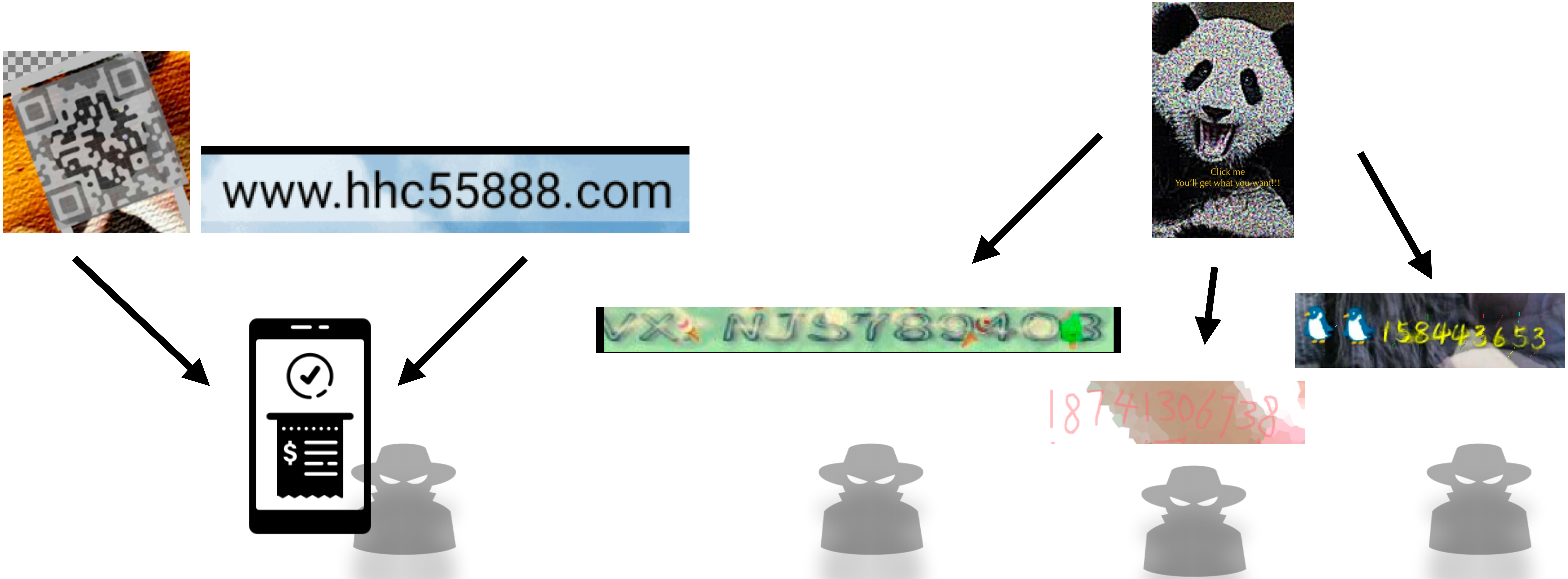








Measurement: Promotional Content



Reuse of promotional content

232 / 612 (37%)

Reuse of explicit content

3981 / 4353 (91%)

MEASUREMENT: PROMOTIONAL CONTENT

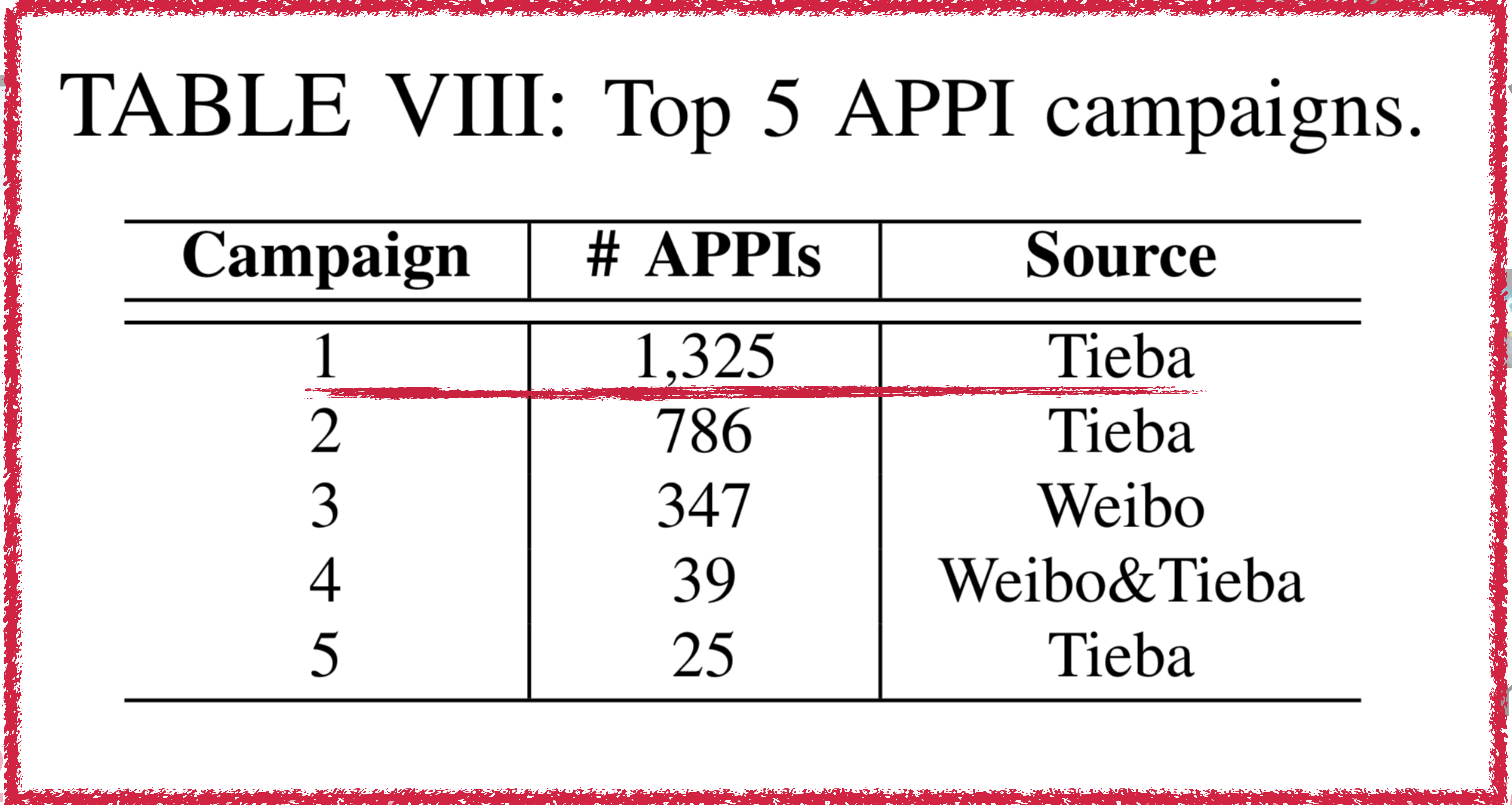
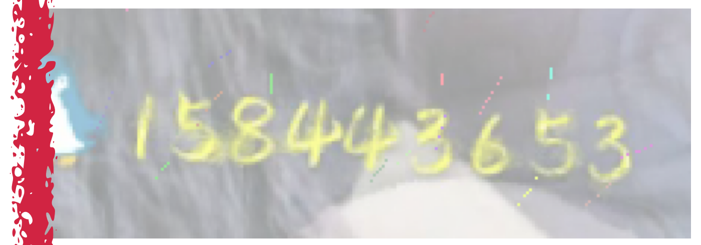


TABLE VIII: Top 5 APPI campaigns.

Campaign	# APPIs	Source
1	1,325	Tieba
2	786	Tieba
3	347	Weibo
4	39	Weibo&Tieba
5	25	Tieba



Reuse

232

ntent

3981 / 4353 (91%)

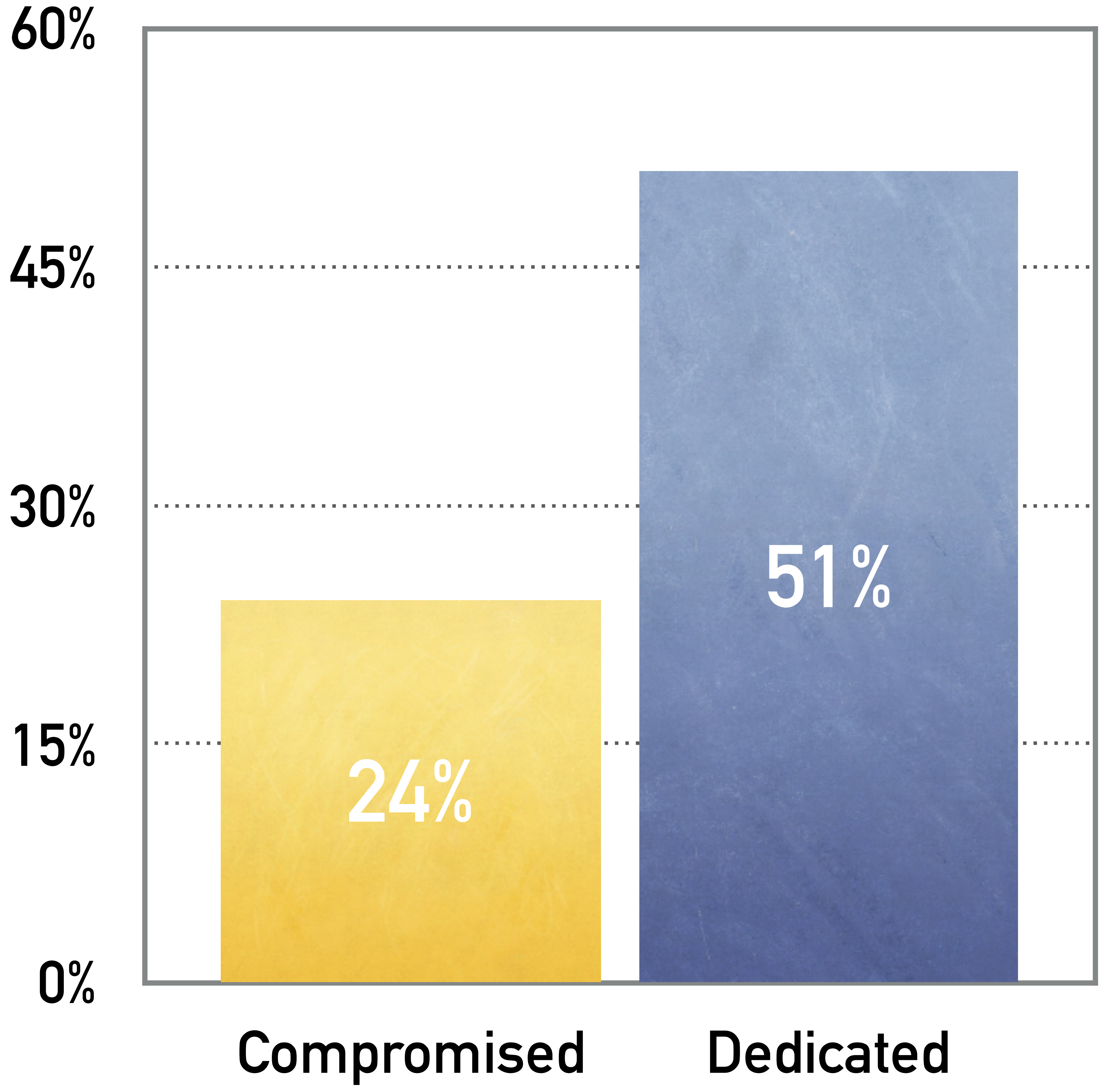
Measurement: Distribution Channels

Compromised accounts:

- rarely post
- comment only on hot microblog

Dedicated accounts:

- > 30 posts/day
- with meaningless sentences



LESSION LEARNED

Visual pattern. → Harden current models.

Promotional content. → Regularize promotion channel.

Distribution channels. → Secure accounts.

Take-aways

- APPIs are prevalent
- Understanding criminal goal and ecosystem behind adversarial images
- Hardening machine learning model against APPI attack deserves further studies

Thank You!