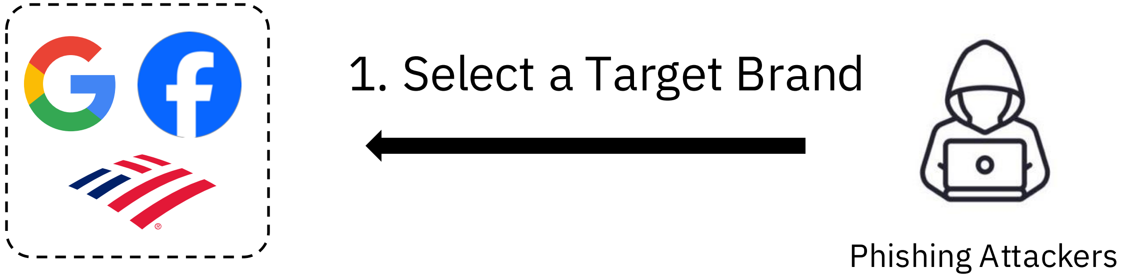
# Evaluating the Effectiveness and Robustness of Visual Similarity-based Phishing Detection Models

Fujiao Ji*, Kiho Lee*, Hyungjoon Koo‡, Wenhao You†, Euijin Choo†, Hyoungshick Kim‡, Doowon Kim*

University of Tennessee, Knoxville*
University of Alberta†
Sungkyunkwan University‡

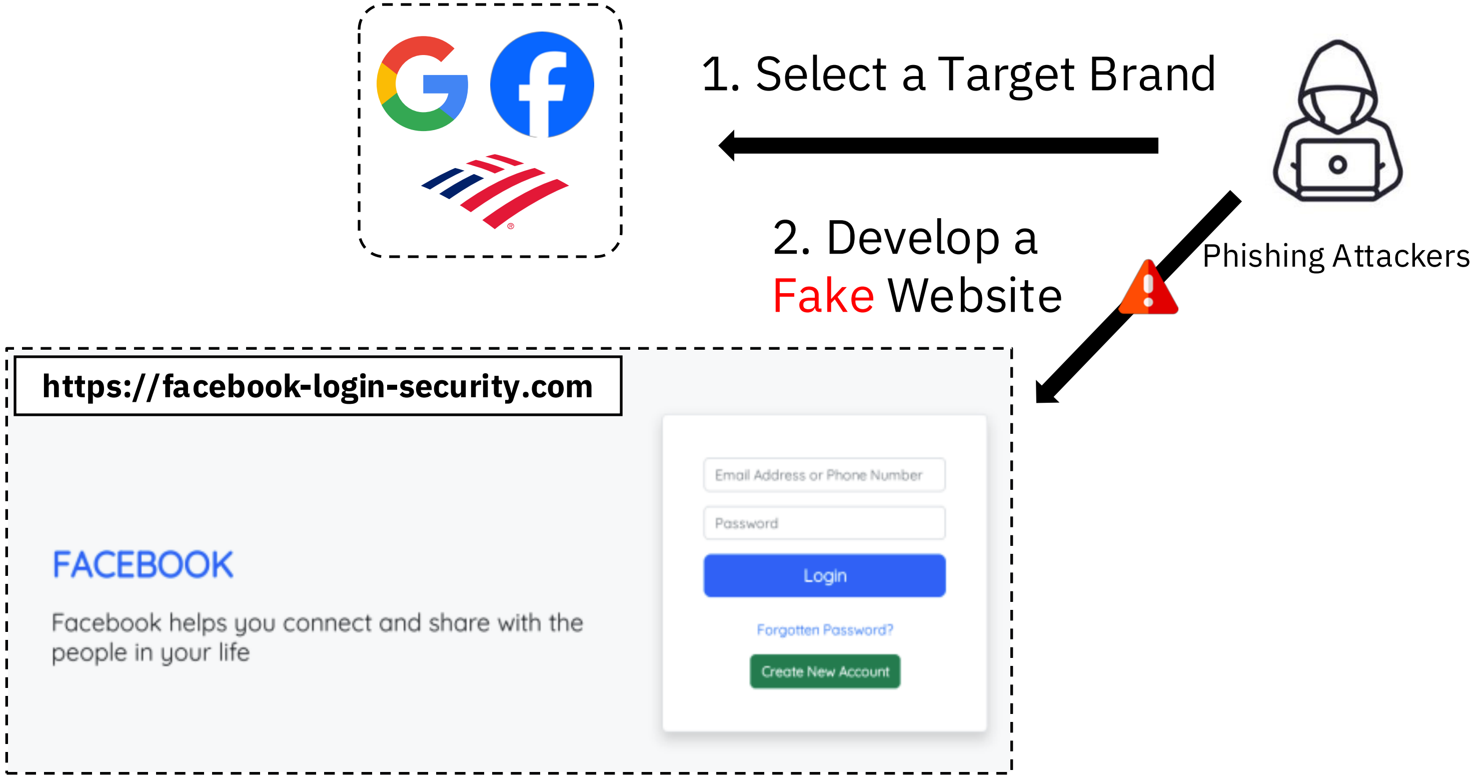
# What Are the Phishing Attacks?



1. Select a Target Brand

Phishing Attackers

# What Are the Phishing Attacks?



1. Select a Target Brand

Phishing Attackers

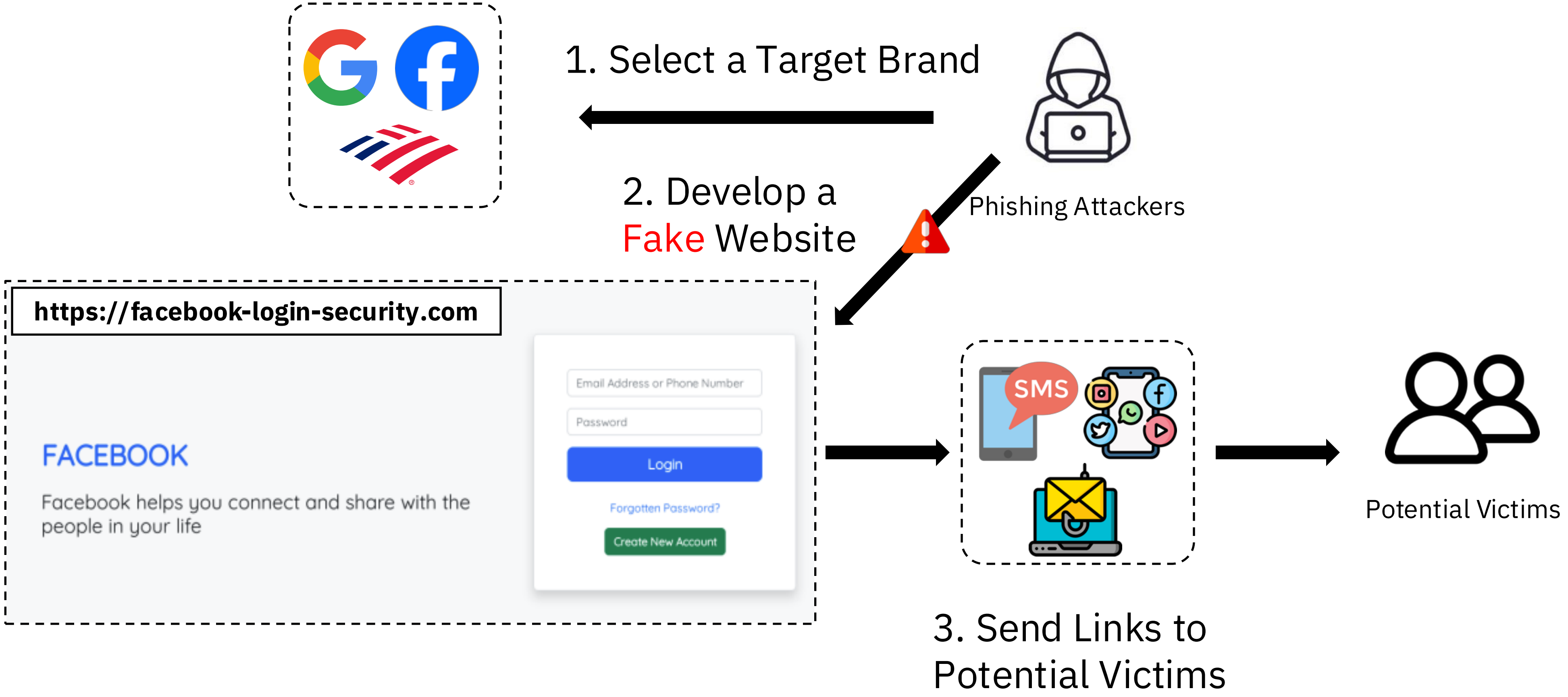
2. Develop a Fake Website

https://facebook-login-security.com

FACEBOOK

Facebook helps you connect and share with the people in your life

# What Are the Phishing Attacks?



1. Select a Target Brand

Phishing Attackers

2. Develop a Fake Website

https://facebook-login-security.com

FACEBOOK

Facebook helps you connect and share with the people in your life

Email Address or Phone Number

Password

Login

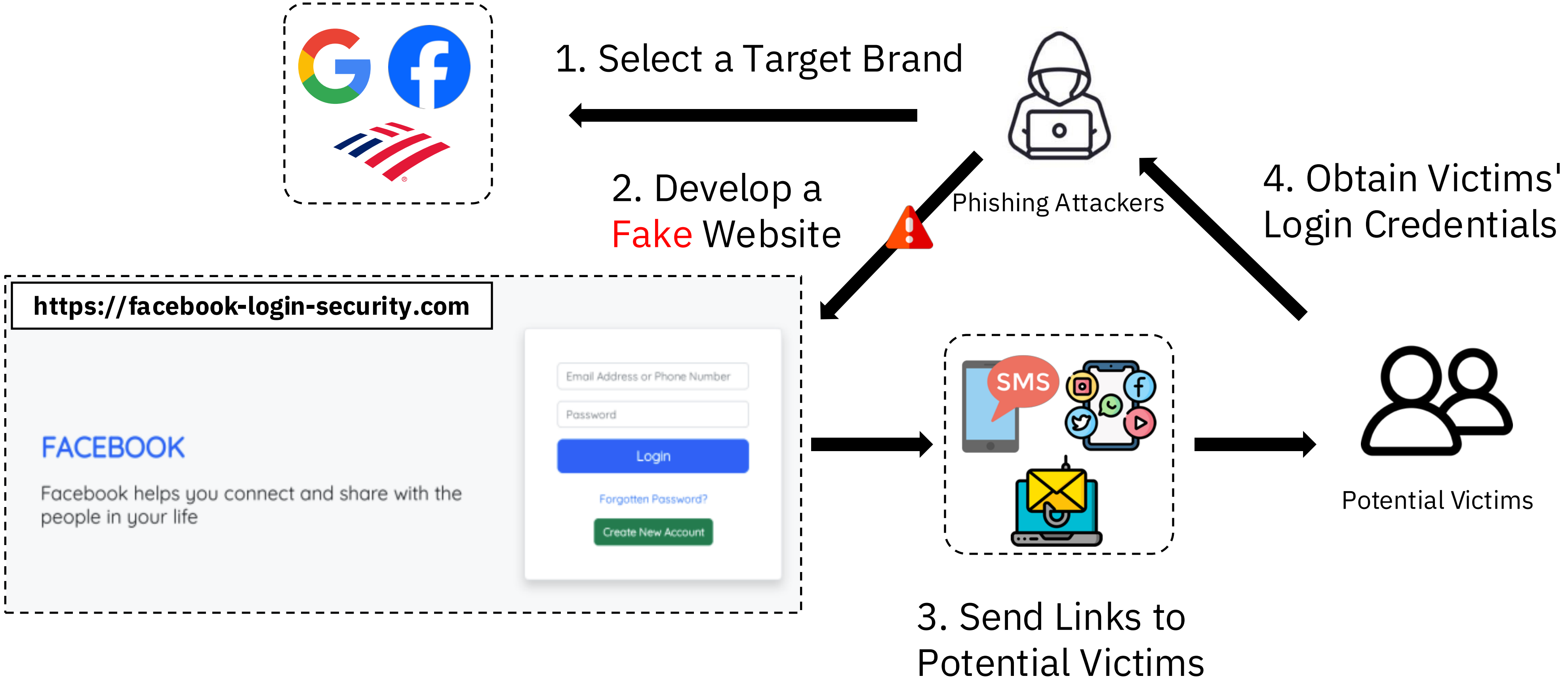
Forgotten Password?

Create New Account

Potential Victims

3. Send Links to Potential Victims

# What Are the Phishing Attacks?



1. Select a Target Brand

2. Develop a Fake Website

3. Send Links to Potential Victims

4. Obtain Victims' Login Credentials

Phishing Attackers

Potential Victims

https://facebook-login-security.com

FACEBOOK

Facebook helps you connect and share with the people in your life

Email Address or Phone Number

Password

Login

Forgotten Password?

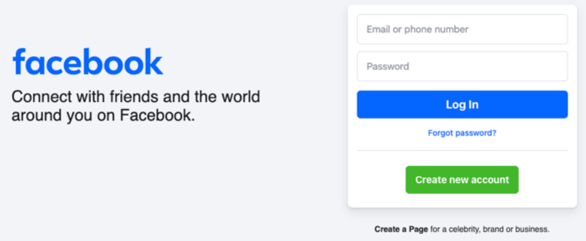
Create New Account

# Current Anti-phishing Systems: Visual Similarity-based Phishing Defense Models

Reference List of Benign Websites

facebook

facebook.com



facebook

Connect with friends and the world around you on Facebook.

Email or phone number

Password

Log In

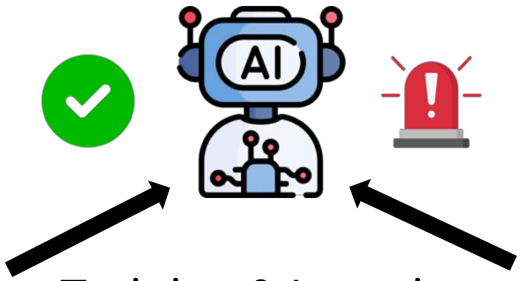
Forgot password?

Create new account

**Create a Page** for a celebrity, brand or business.

Training & Learning

# Current Anti-phishing Systems:
# Visual Similarity-based Phishing Defense Models

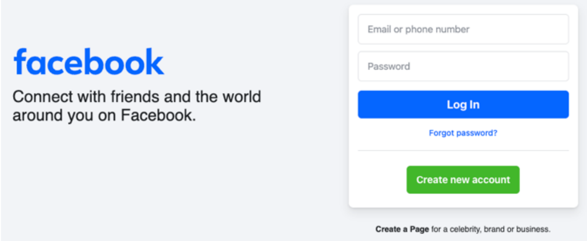


Training & Learning

Reference List of Benign Websites

facebook

facebook.com



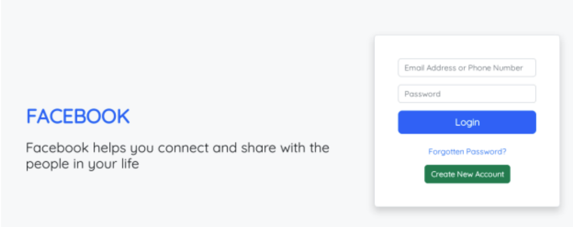
facebook

Connect with friends and the world around you on Facebook.

Potential Phishing Websites

FACEBOOK

faceb00k.com



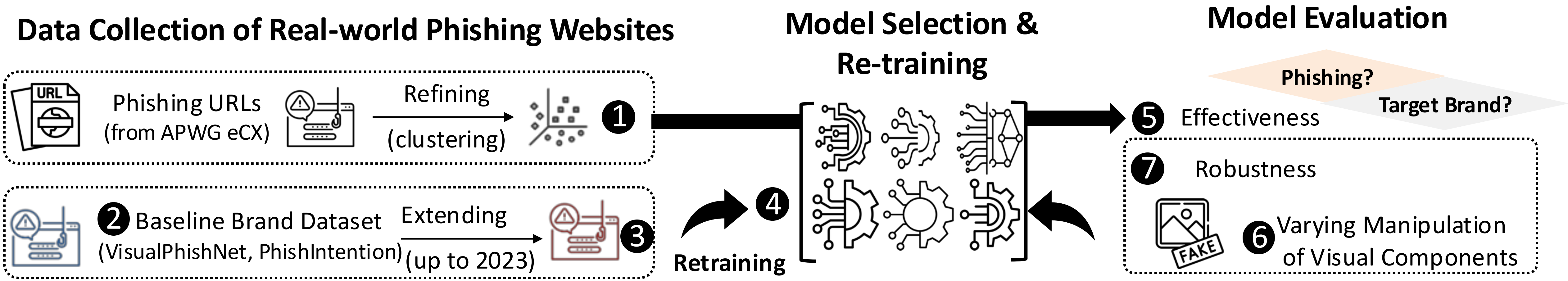
FACEBOOK

Facebook helps you connect and share with the people in your life

# Main Research Question

Are these current phishing detection models (visual similarity-based) **effective** against real-world phishing websites and **robust** to adversarial strategies?

# Our Evaluation Pipeline



**Data Collection of Real-world Phishing Websites**

Phishing URLs (from APWG eCX) → Refining (clustering) ❶

❷ Baseline Brand Dataset (VisualPhishNet, PhishIntention) → Extending (up to 2023) ❸

**Model Selection & Re-training**

❹ **Retraining**

**Model Evaluation**

❺ Effectiveness

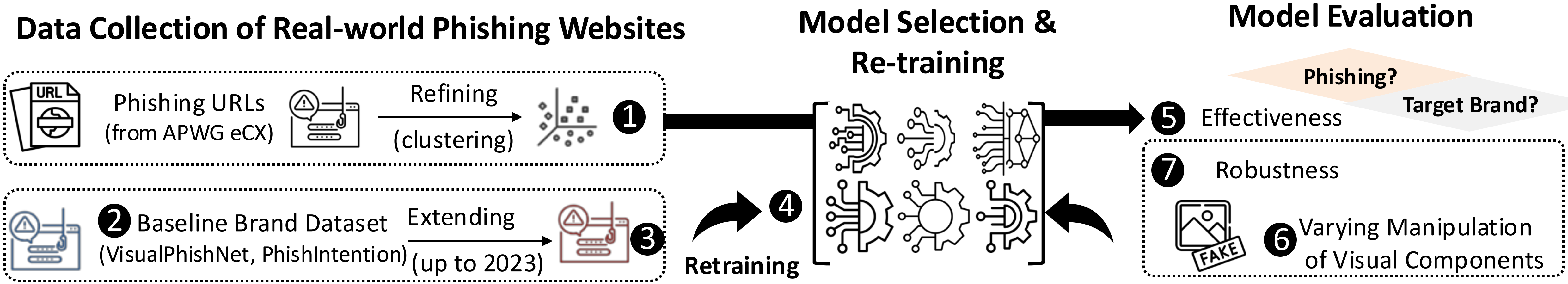
**Phishing?**

**Target Brand?**

❼ Robustness

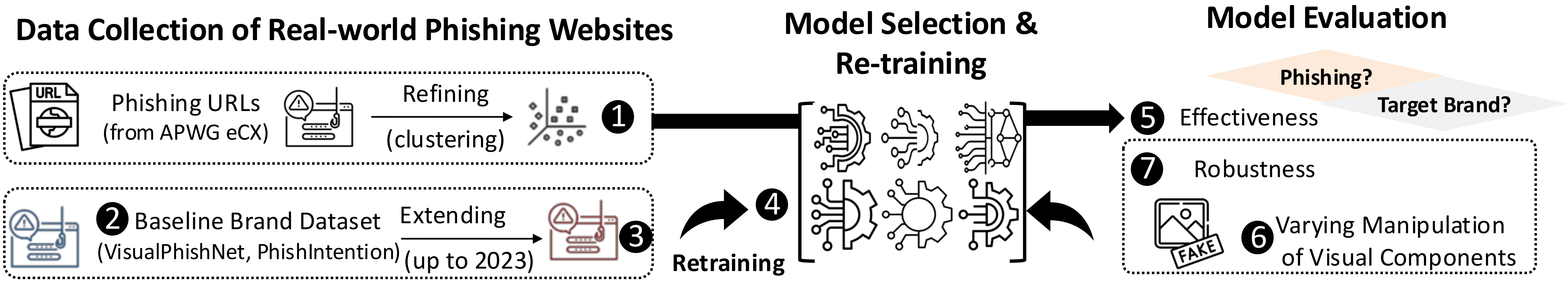
❻ Varying Manipulation of Visual Components

# Our Evaluation Pipeline



- Developed a web-crawler that visits phishing websites fed by APWG
- Collected from July 2021 to July 2023 (**25 months**) → **6.1M** samples
- Obtained **451k** samples after removing error pages and sampling

# Our Evaluation Pipeline



Data Collection of Real-world Phishing Websites

Model Selection & Re-training

Model Evaluation

Phishing URLs (from APWG eCX) Refining (clustering) ❶

❷ Baseline Brand Dataset (VisualPhishNet, PhishIntention) Extending (up to 2023) ❸

Retraining ❹

❺ Effectiveness Phishing? Target Brand?

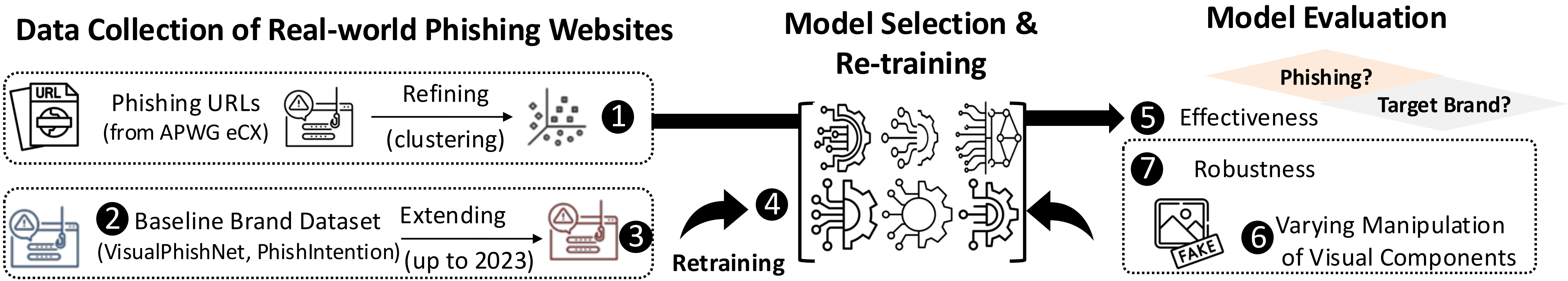
❼ Robustness

❻ Varying Manipulation of Visual Components

| | |
|---|---|
| PhishIntention | PhishZoo |
| Phishpedia | VisualPhishNet |
| DynaPhish | EMD |
| Involution | |

***Retraining***: To ensure fair evaluation, the models should share the same reference knowledge of brands.

# Our Evaluation Pipeline



Using a real-world phishing dataset and a manipulated dataset to evaluate effectiveness and robustness.

# Results: Overall Detection Performance

- Detection performance **degradation** (**20.7%**) compared to their results on curated datasets

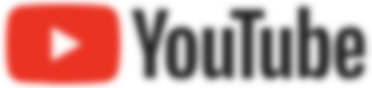
| Models | Ref. Type | **Detection** ($R_{ext}$) | | | **Identification** ($R_{ext}$) | |
|---|---|---|---|---|---|---|
| | | $N_{tp}$ for $D_{all}$ ($N_p$: 451,514) | $N_{tp}$ for $D_{learn}$ ($N_p$: 312,355) | $N_{tp}$ for $D_{sample}$ ($N_p$: 4,190) | $D_{learn}$ $I_{tp}/N_{tp}$ | $D_{sample}$ $I_{tp}/N_{tp}$ |
| DynaPhish | Logo | ---- | ---- | 22.03% | ---- | 97.94% |
| PhishIntention | Logo | 52.68% | 66.22% | 49.07% | 97.72% | 98.56% |
| Phishpedia | Logo | 70.47% | 87.97% | 57.16% | 96.67% | 92.36% |
| Involution | Logo | 66.67% | 84.77% | 60.57% | 99.64% | 97.32% |
| PhishZoo | Logo | 86.28% | 86.36% | 76.13% | 33.26% | 9.59% |
| VisualPhishNet | Scr. | 41.33% | 40.58% | 33.84% | 66.03% | 54.51% |
| EMD | Scr. | 30.28% | 31.34% | 27.45% | 22.91% | 20.42% |

# Detection-Failed Cases (Three Adversarial Strategies)

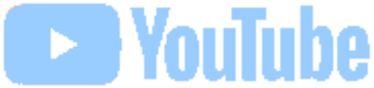
1) Model Pipeline Attack

Benign

Phishing



Blurred

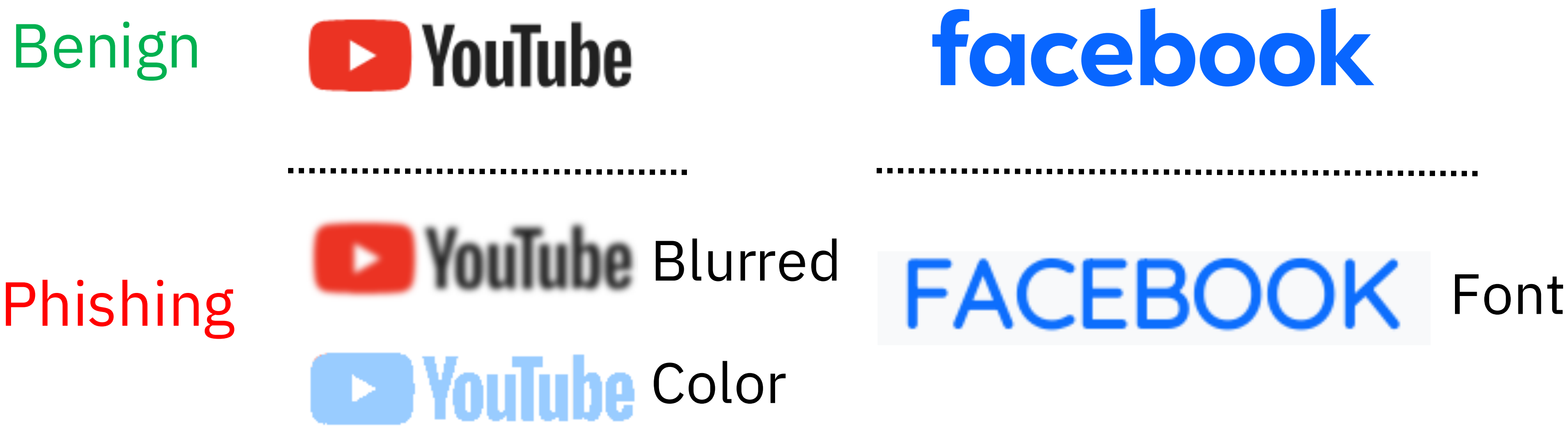


Color

# Detection-Failed Cases (Three Adversarial Strategies)

1) Model Pipeline Attack

2) Mimic Visualization



Benign

Phishing

Blurred

Color

Font

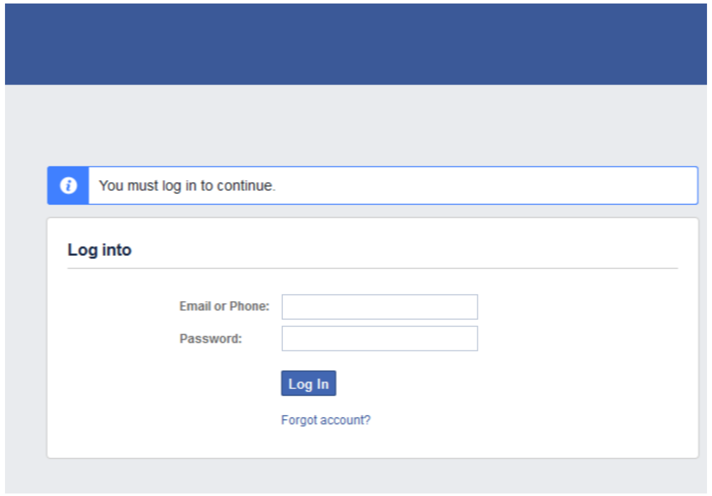
# Detection-Failed Cases (Three Adversarial Strategies)

1) Model Pipeline Attack

2) Mimic Visualization

3) Direct Simple Strategies

Benign


You must log in to continue.

Log into

Email or Phone:

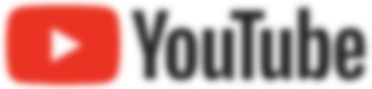
Password:

Log In

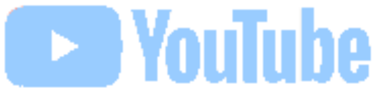
Forgot account?

Logo Elimination

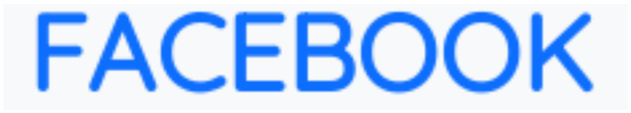
Phishing



Blurred



Color



Font

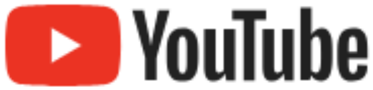
THE UNIVERSITY OF TENNESSEE KNOXVILLE

# Robustness of Visible and Perturbation-based Manipulations

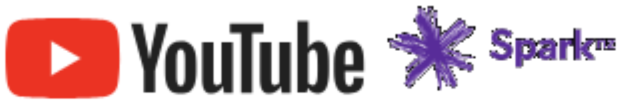
# Robustness of Visible and Perturbation-based Manipulations

- **Logo-based methods** are disrupted for brand identification (Phishpedia: 15.72% for integration, 16% for case conversion);
- **Screenshot-based methods** exhibit lower detection rate (VisualPhishNet: 27.27% on benign samples).

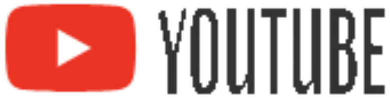
Original



Integration



Case Conversion



# Key Takeaways

1. Performance degradation (**20.7%**) compared to their results on curated datasets
2. Need for robust, multi-modal defenses that don't overly rely on single features (e.g., logos or exact visual patterns)
3. The dataset is publicly available at https://moa-lab.net/evaluation-visual-similarity-based-phishing-detection-models/