

# **PinkDroid**

## **A System Level Contents Blocker**

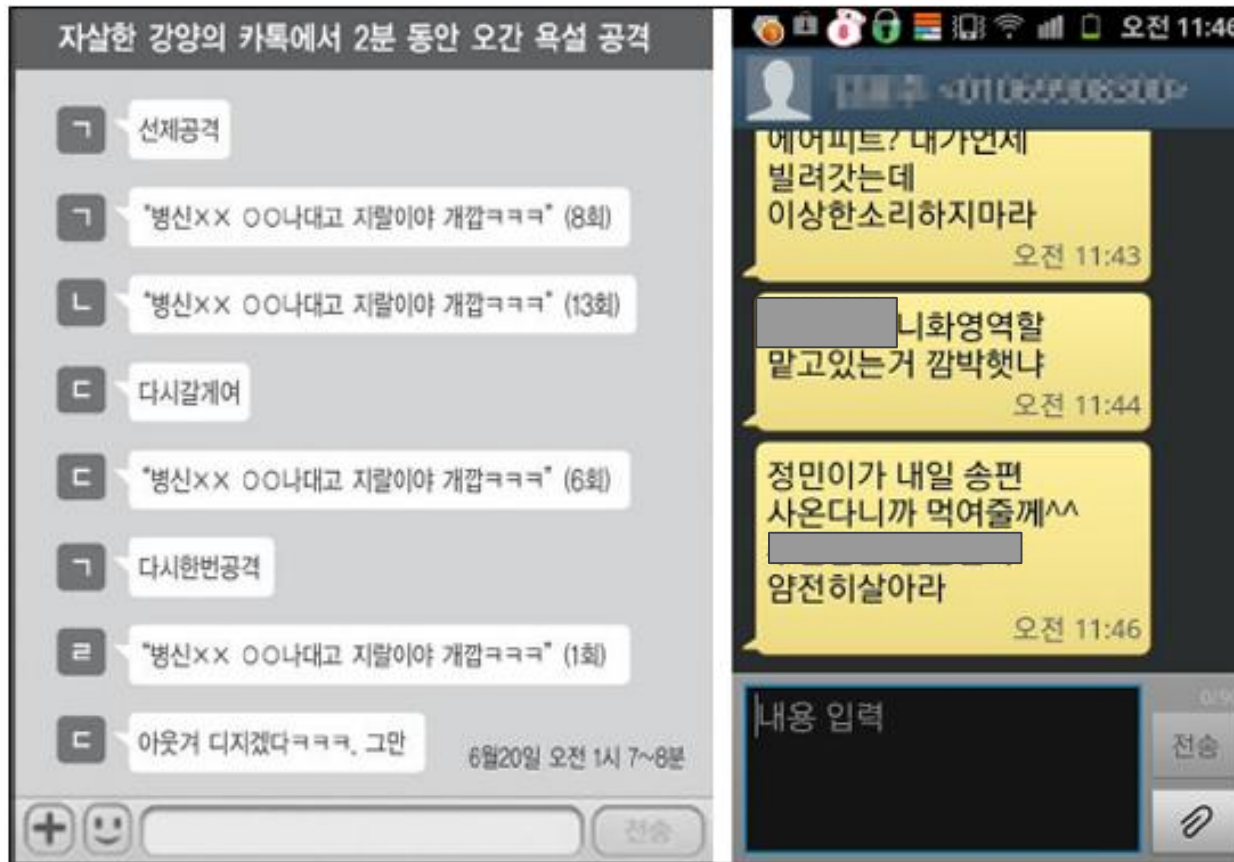
**Wonsup Yoon, Bilgehan Bingol, Nak Hyun Choi**

A group of seven diverse children are posed on a grassy field. In the back row, a boy with short brown hair in a maroon hoodie, a girl with long blonde hair in a blue top, and a boy with curly brown hair in a blue shirt. In the middle row, a girl with long dark hair in a red jacket and a boy with blonde hair in a blue and white striped shirt. In the front row, a girl with long dark hair in a light beige jacket and a girl with large dark curly hair in a dark blue jacket over an orange shirt. All children are smiling and looking towards the camera. The background is a soft-focus green field with trees.

Children are  
important!

# But in their smartphones...

---



## Bad words



But in their smartphones...



Adult contents

But in their smartphones...

— — —



Violence

But in their smartphones...

— — —

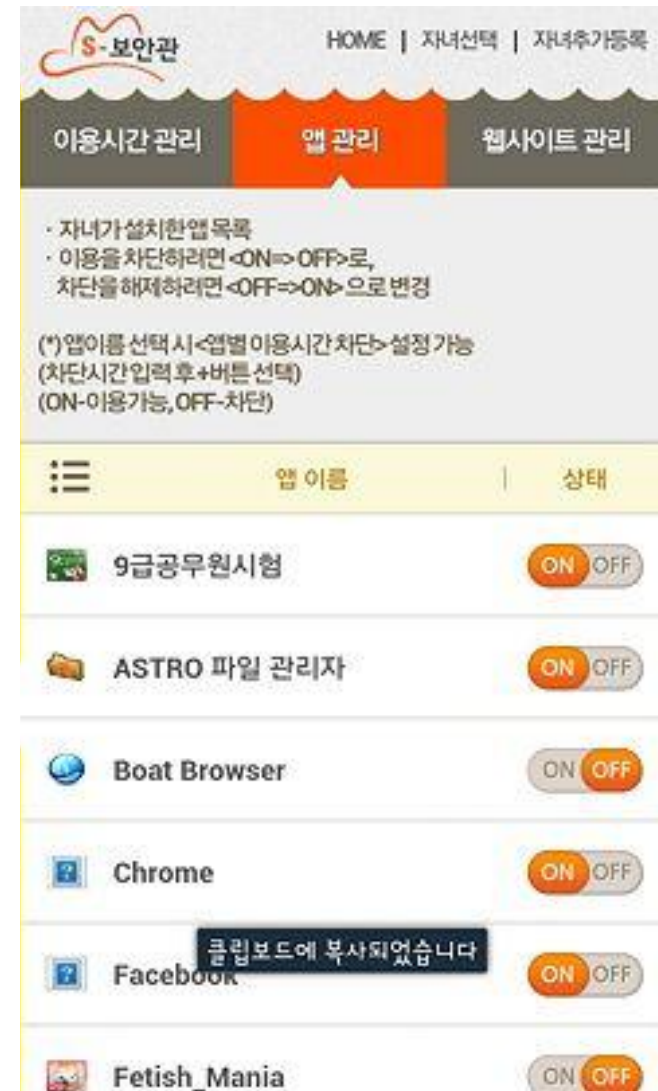


Medical

**We must protect our children from the  
inappropriate contents!**

# Previous solution

- Only blocks applications (block or permit)
- Is it worthwhile to block KakaoTalk or Facebook?
- Even in the permitted applications, there can be inappropriate contents

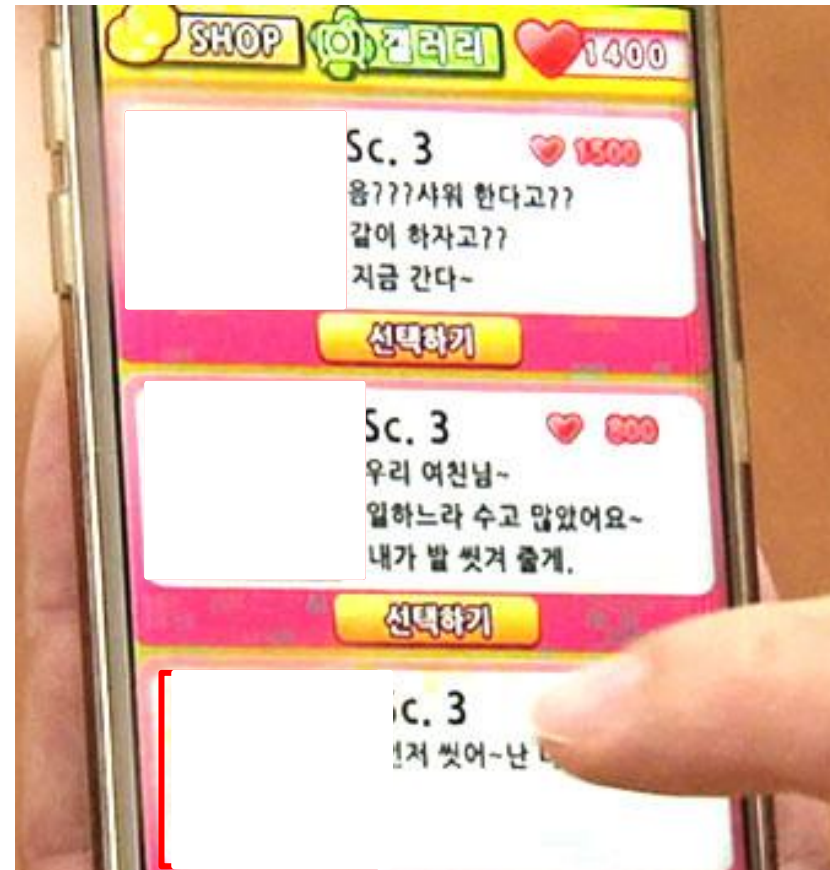




# Proposed Solution

---

- Block inappropriate contents at the UI rendering
- Detect which contents(text, images) are inappropriate
- Block that contents at runtime



# Proposed Solution

---

- Anti-Social Dictionary
- Content Trust Model
- External API

# Anti-Social Dictionary

---

- Target contents: Texts
- Using predefined text set
- Simple string matching
- But, it is not suitable for photos.

# Content Trust Model

---

- Target contents: Photos
- Tracking all photos in a smartphone
- Determine drawable photos with their sources or characteristics.



# Content Trust Model

Trusted	Untrusted
<ul style="list-style-type: none"><li>● Built with firmware image</li><li>● Built with trusted apps *</li><li>● User generated contents ** (ex. camera)</li><li>● Contents generated from trusted contents ***</li><li>● Small images ****</li><li>● ...</li></ul>	<ul style="list-style-type: none"><li>● Downloaded from the Internet</li><li>● Built with untrusted apps</li><li>● Sended from outside</li><li>● Contents generated from untrusted contents ***</li><li>● ...</li></ul>

\* If Google Play testers work very hard and don't make mistakes

\*\* If your children are not anti-social

\*\*\* Trusts can be inherited

\*\*\*\* If a photo is small enough, it is unable to store much information. It should be used for UI.

# Content Trust Model

Trusted	Untrusted
<ul style="list-style-type: none"><li>● <b>Built with firmware image</b></li><li>● <b>Built with trusted apps *</b></li><li>● User generated contents ** (ex. camera)</li><li>● <b>Contents generated from trusted contents ***</b></li><li>● <b>Small images ****</b></li><li>● ...</li></ul>	<ul style="list-style-type: none"><li>● Downloaded from the Internet</li><li>● Built with untrusted apps</li><li>● Sended from outside</li><li>● Contents generated from untrusted contents ***</li><li>● ...</li></ul>

\* If Google Play testers work very hard and don't make mistakes

\*\* If your children are not anti-social

\*\*\* Trusts can be inherited

\*\*\*\* If a photo is small enough, it is unable to store much information. It should be used for UI.

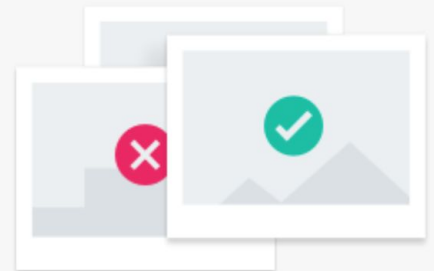
# External API

---

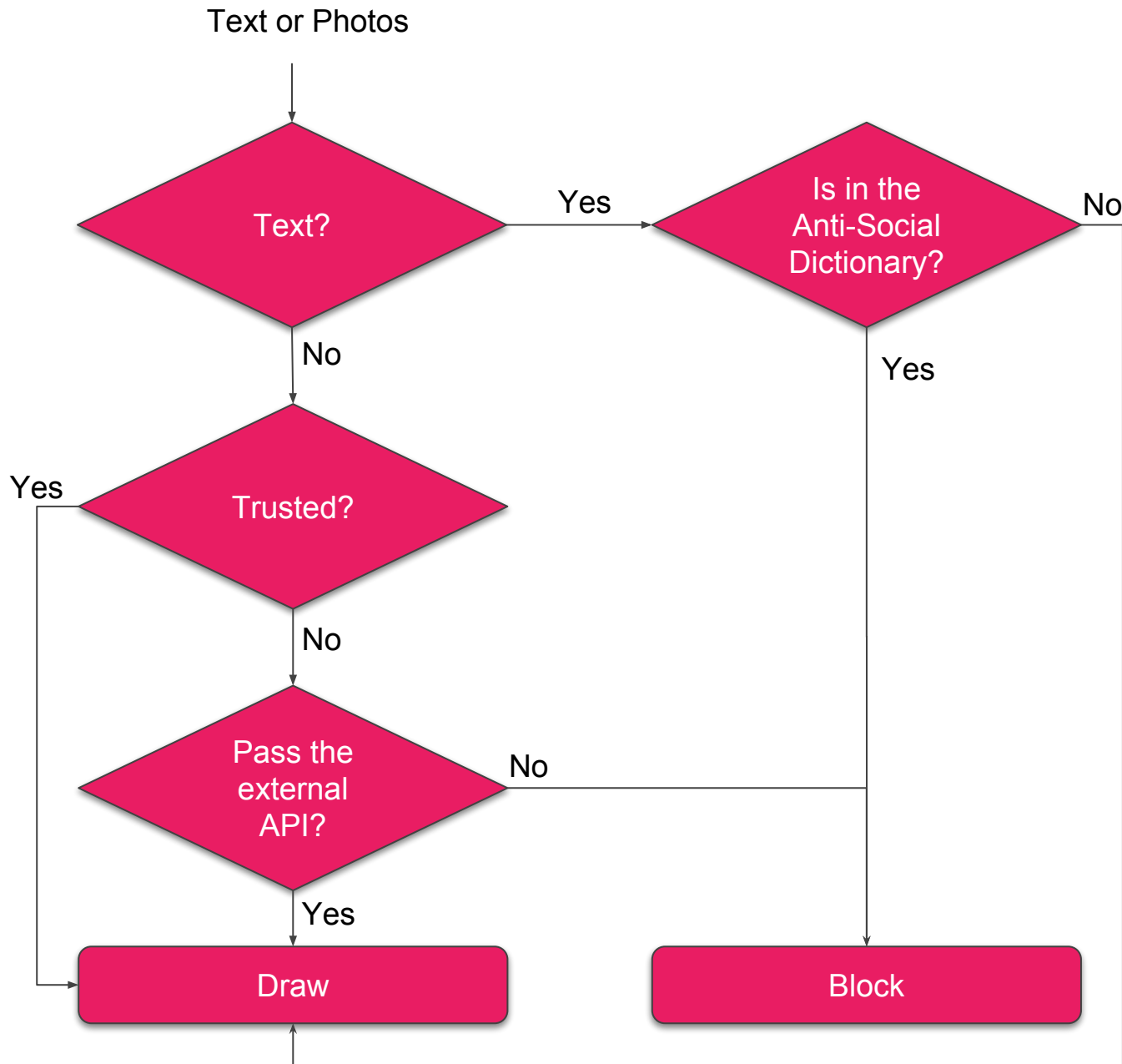
- Target contents: Untrusted photos
- Google offers Cloud Vision API
  - It can detect spoof, medical, adult, and violence contents.
- It can be used via web API.
- For performance reasons results should be cached.
- It can be replaced by any other Deep learning frameworks.

## Detect Inappropriate Content

Powered by Google [SafeSearch](#), **easily moderate content** from your crowd sourced images. Vision API enables you to detect different types of inappropriate content from adult to violent content.



\* <https://cloud.google.com/vision/>





# Evaluation

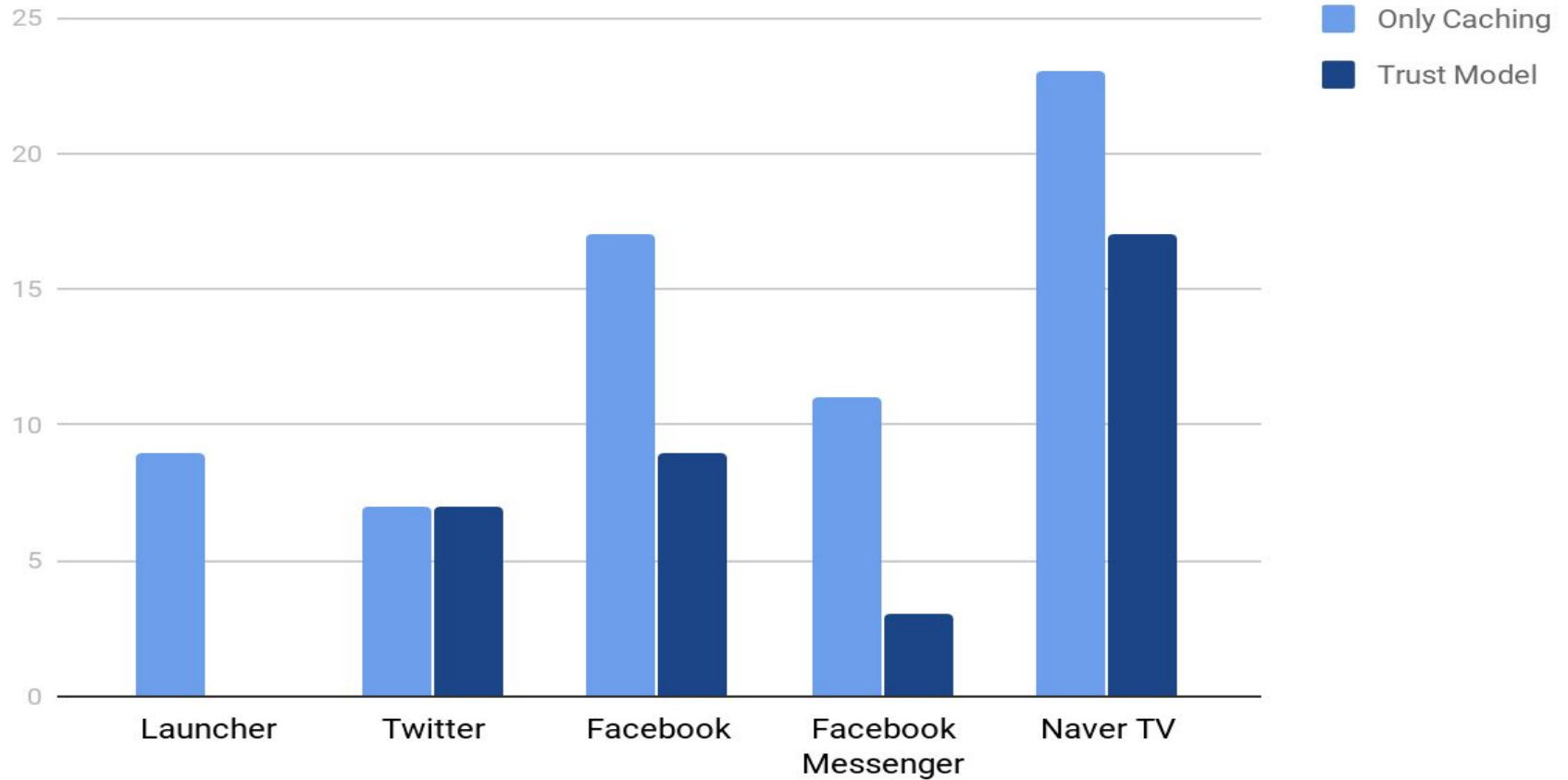
# Evaluation

---

- Evaluation
  - Performance (Cost): # of images sent to Google
    - \$1.50 for 1,000 images
  - Safety: # of images filtered
- Environment
  - Android 7.1.1
  - aosp\_x86-eng
  - Testing Apps
    - Launcher
    - Twitter
    - Facebook
    - Facebook Messenger
    - Naver TV

# Performance evaluation (Only large images)

# of requests

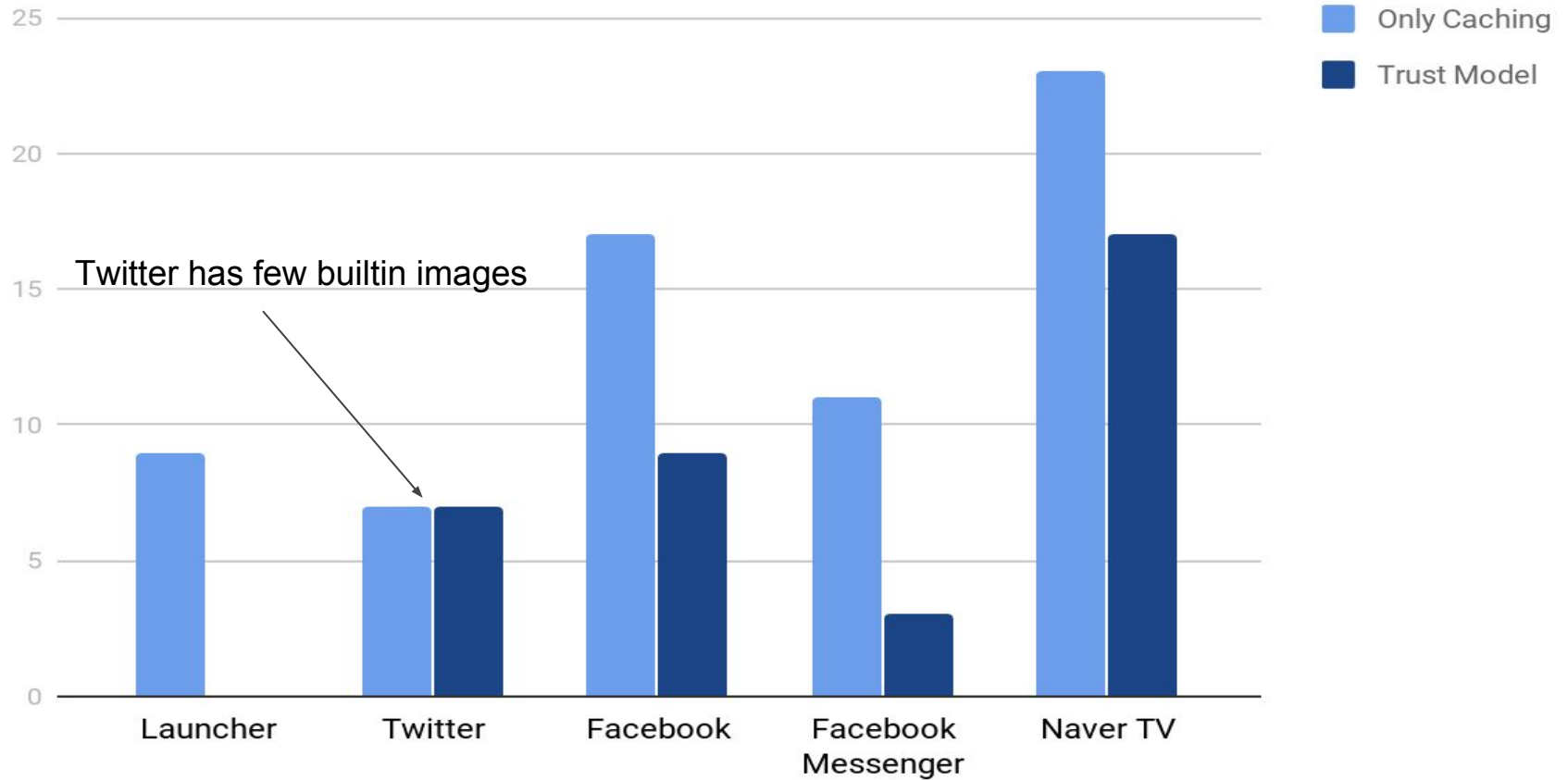


*Lower is Better*

\* Workloads are not exactly same.

# Performance evaluation (Only large images)

# of requests



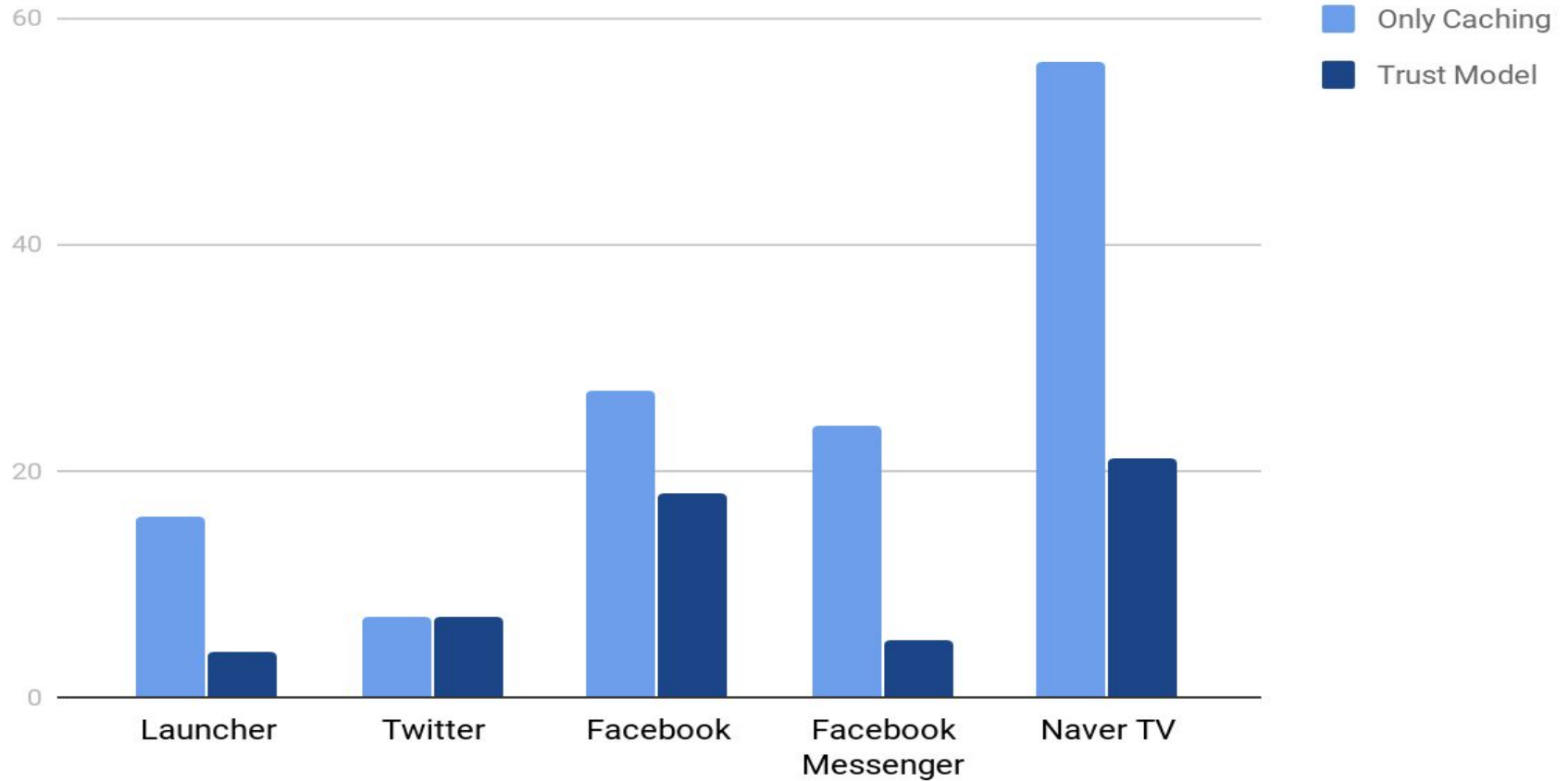
*Lower is Better*

\* Workloads are not exactly same.



# Performance evaluation (All images)

# of requests

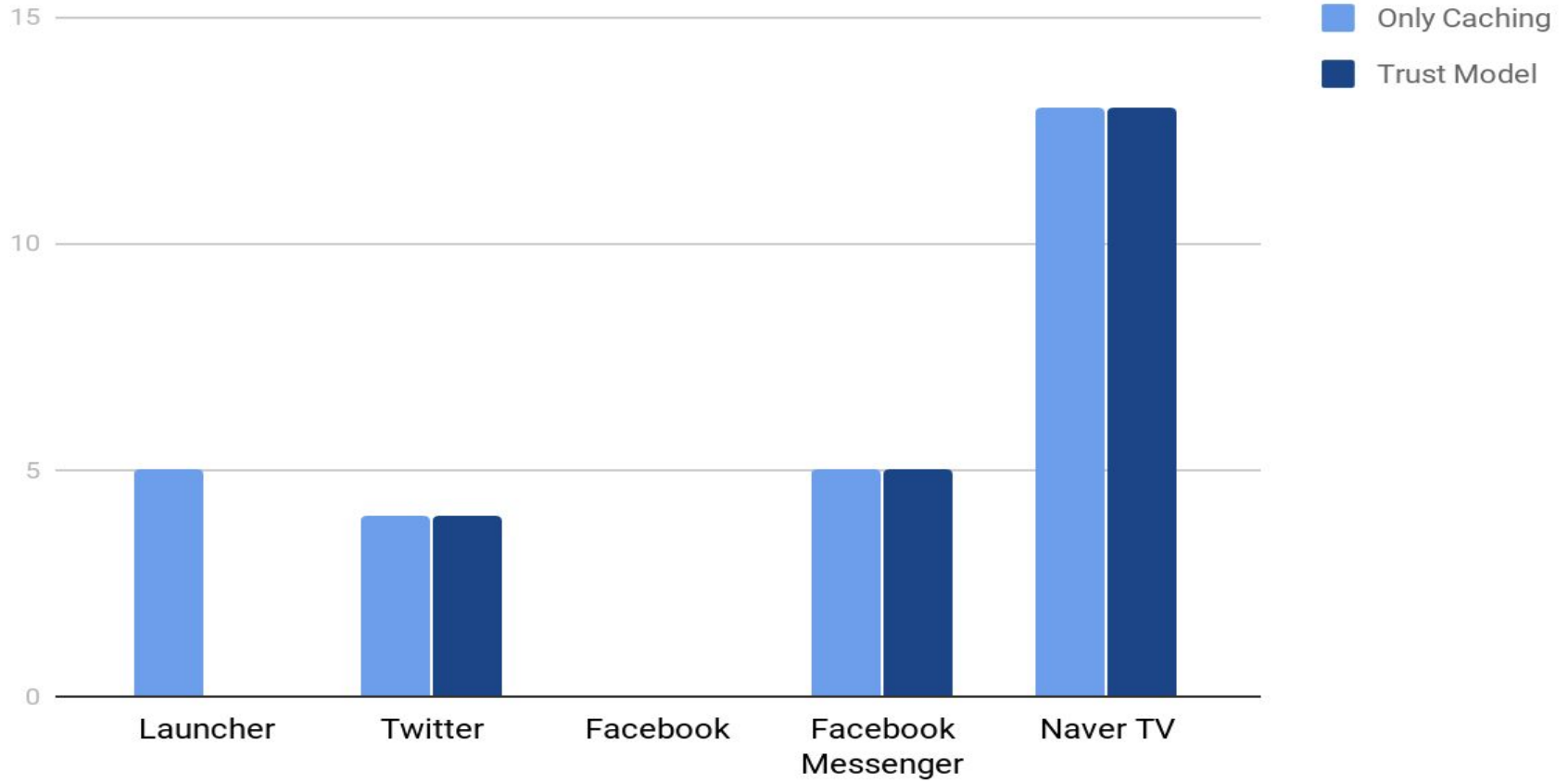


*Lower is Better*

\* Workloads are not exactly same.

# Safety evaluation

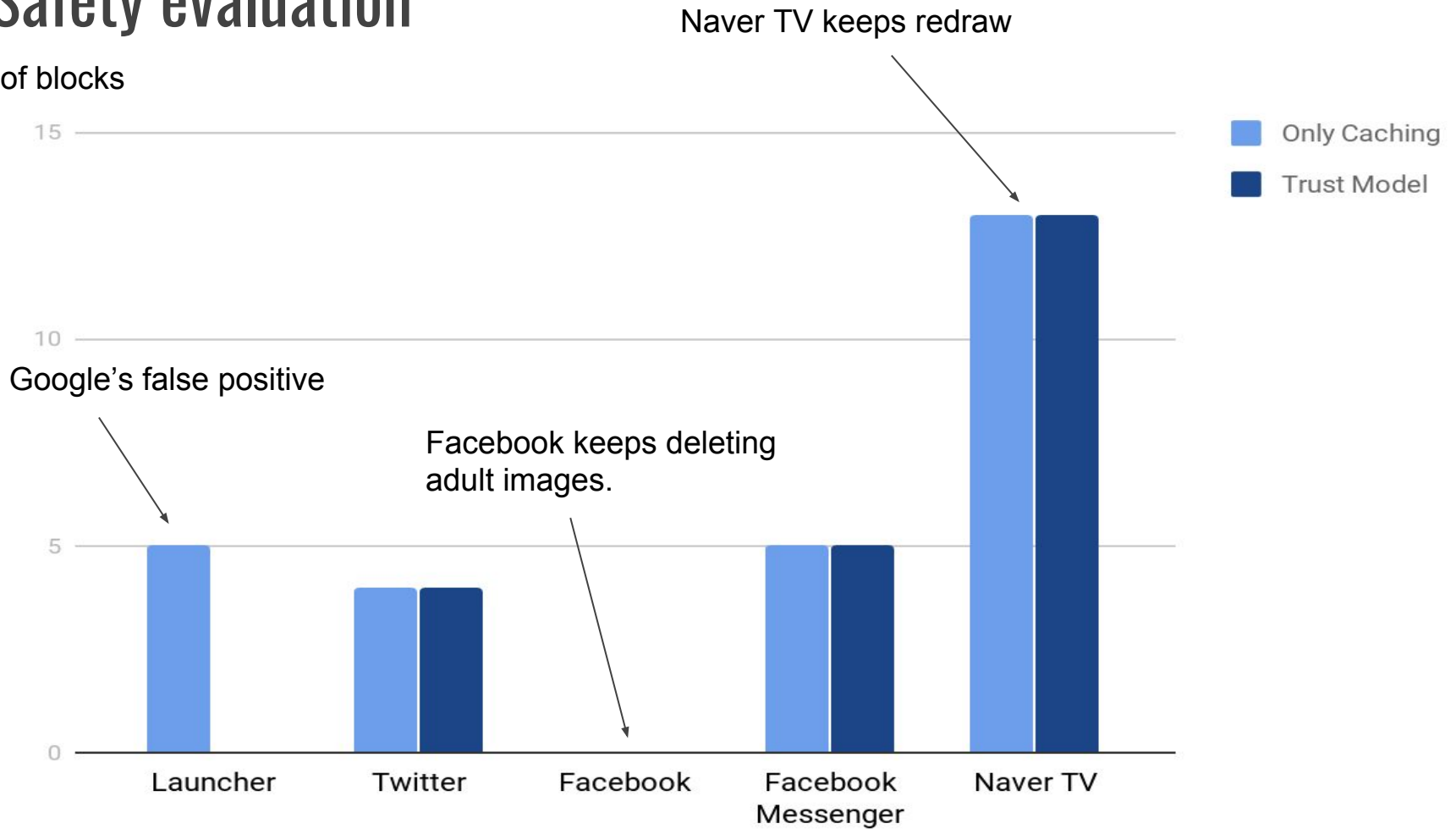
# of blocks



\* Workloads are not exactly same.

# Safety evaluation

# of blocks



\* Workloads are not exactly same.

*Demo*

# Facebook & Facebook Messenger

# Actors

---



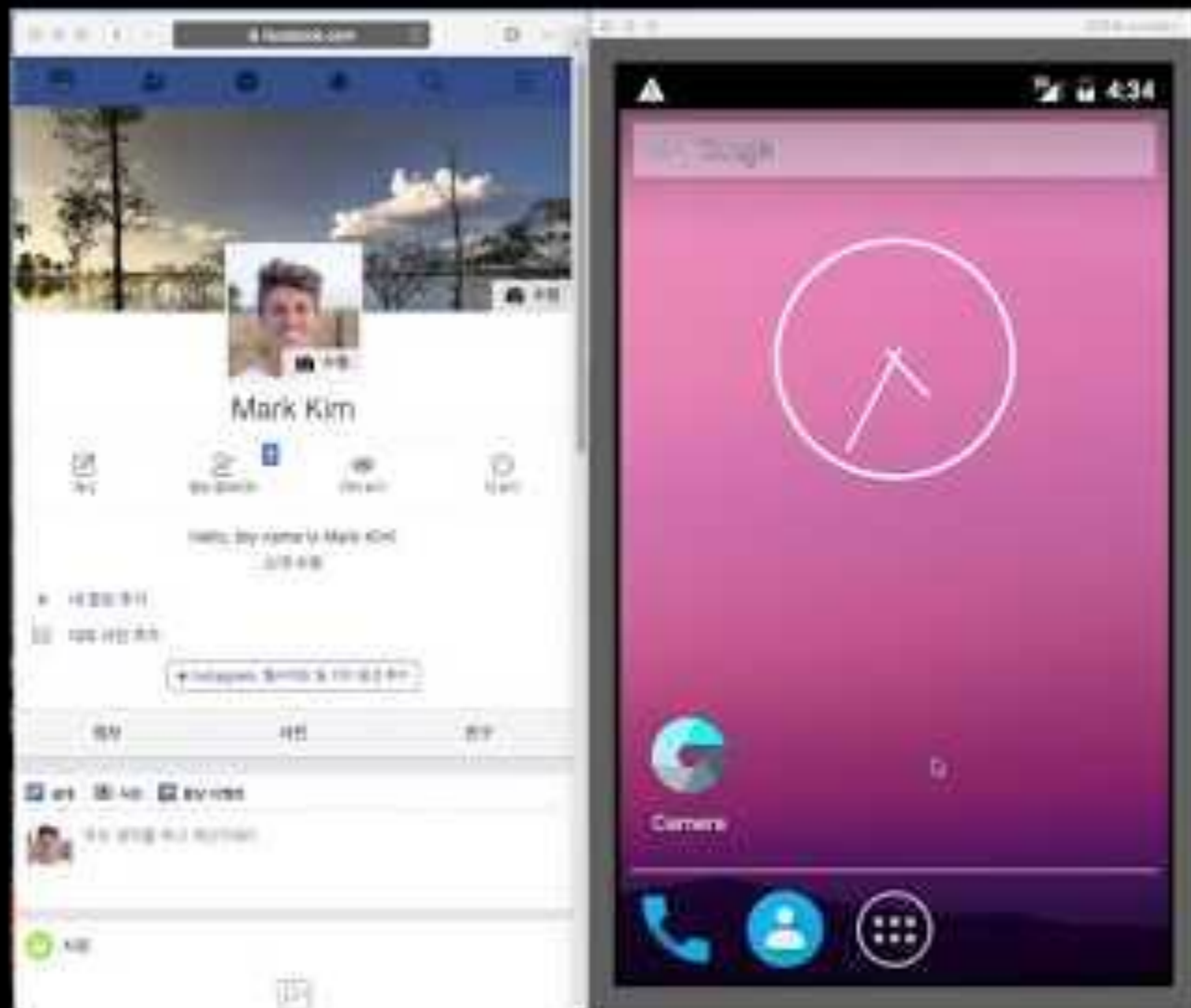
Wonsup Yoon

A Tiny Little Boy



Mark Kim

Creepy Pedophile  
Sex Offender





**Naver TV**

NAVER TV

홈 피드 TOP100 라이브 TV방송 테마

가장 인기 있는 콘텐츠

가장 인기 있는 콘텐츠

- 1 [최초 공개] 브라운보이즈 - '점거부' @X-CON  
Wanna One(워너원) GO  
11.22일 10.52%
- 2 수식어가 필요 없는 박정현 'Someone Like You'  
박정현(박정현) GO  
11.22일 10.52%
- 3 [최초 공개] 안원미 - '영원+1' @X-CON  
Wanna One(워너원) GO  
11.22일 10.52%
- 4 한승연&조문, 김은 장미에서 는 프다 (X-CON)  
Wanna One(워너원) GO  
11.22일 10.52%
- 5 김경수의 직전 고백 " 좋아하니까, 다 할고 싶습니다"  
Wanna One(워너원) GO  
11.22일 10.52%

NAVER TV

TOP100 피드 추천 라이브

가장 인기 있는 콘텐츠

가장 인기 있는 콘텐츠

- 1 [최초 공개] 브라운보이즈 - '점거부' @X-CON  
Wanna One(워너원) GO  
11.22일 10.52%
- 2 수식어가 필요 없는 박정현 'Someone Like You'  
박정현(박정현) GO  
11.22일 10.52%
- 3 [최초 공개] 안원미 - '영원+1' @X-CON  
Wanna One(워너원) GO  
11.22일 10.52%

**Thank You!**

