ROME: Testing Image Captioning Systems via Recursive Object Melting

Boxi Yu
CHUKSZ, China

Zhiqing Zhong
CHUKSZ, China

Jiaqi Li
CHUKSZ, China

Yixing Yang
CHUKSZ, China

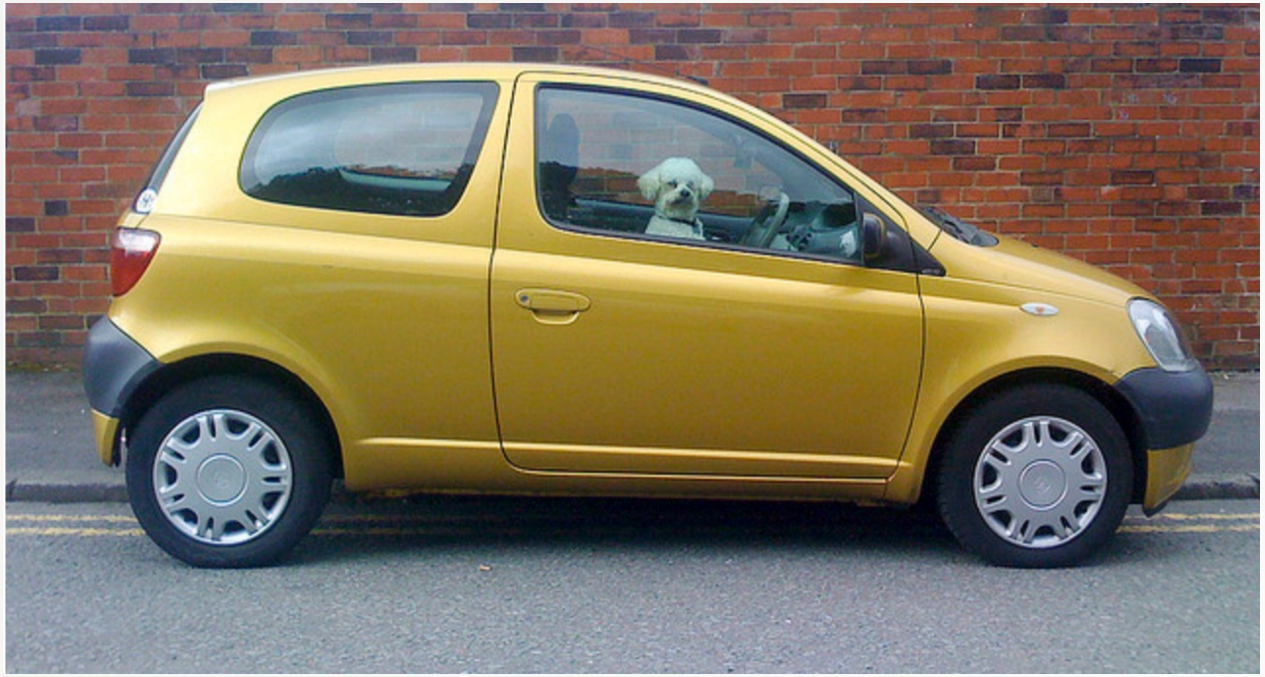
Shilin He
Microsoft, China

Pinjia He
CHUKSZ, China

香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

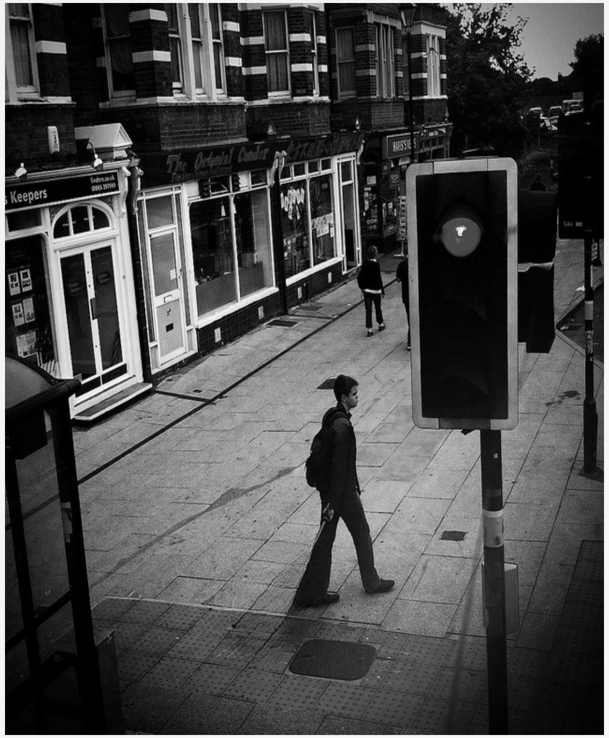
数据科学学院
School of Data Science

Image Caption: Translating Image into Textual Description

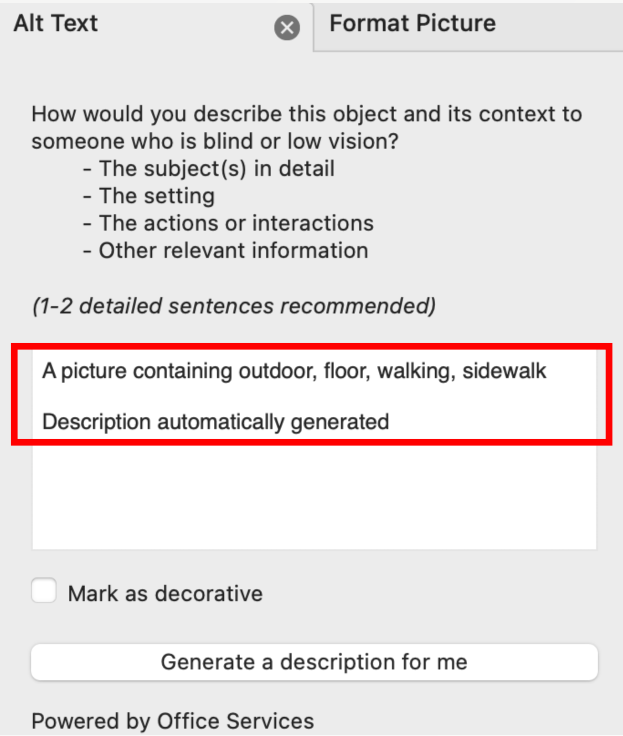


a car with a dog inside it parked in the street.

Real-world IC Software in MS Powerpoint



Input Image



Alt Text ⊗ Format Picture

How would you describe this object and its context to someone who is blind or low vision?

- The subject(s) in detail
- The setting
- The actions or interactions
- Other relevant information

(1-2 detailed sentences recommended)

A picture containing outdoor, floor, walking, sidewalk

Description automatically generated

□ Mark as decorative

Generate a description for me

Powered by Office Services

Microsoft Powerpoint ALT Text

ROME

Recursive **O**bject **Me**lting

Idea: The composition of **objects** within an image should encompass the **objects** derived from its generated descendant images through the process of **recursive object melting**

ROME

Remove the **cat**

Ancestor: A **zebra** standing in the field with a **cat**

Descendent: A **zebra** standing in the field

ROME

Captions

Ancestor Caption: A **zebra** standing in the field with a **cat**

Descendent Caption: A **zebra** standing in the field

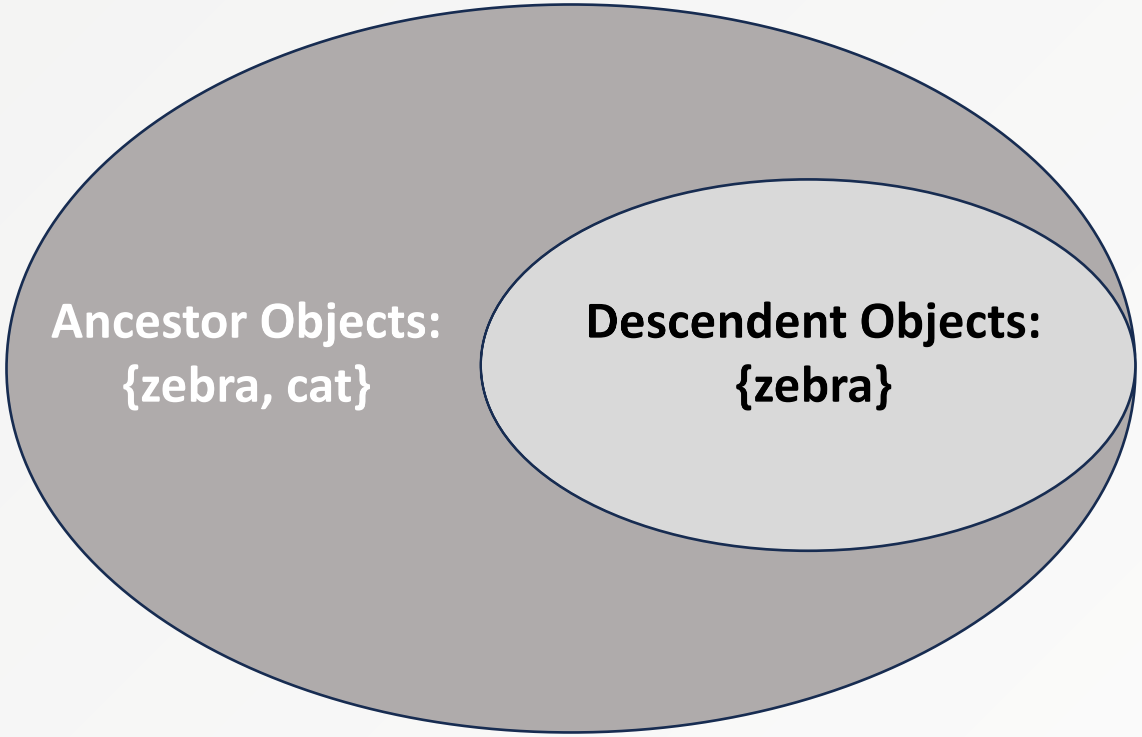
Part-of-speech Tagging

Object Sets

Ancestor objects: {**zebra, cat**}

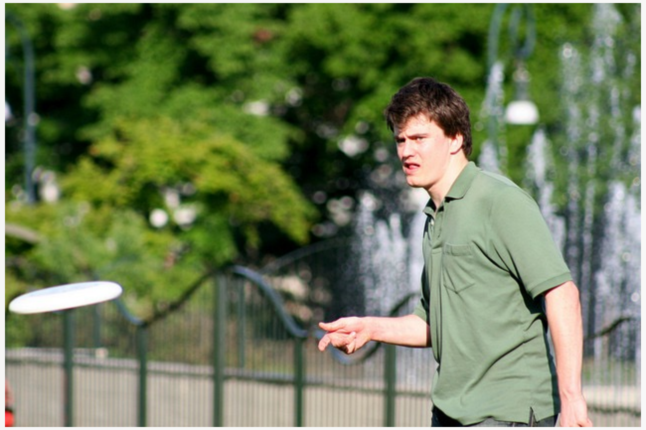
Descendent objects: {**zebra**}

ROME

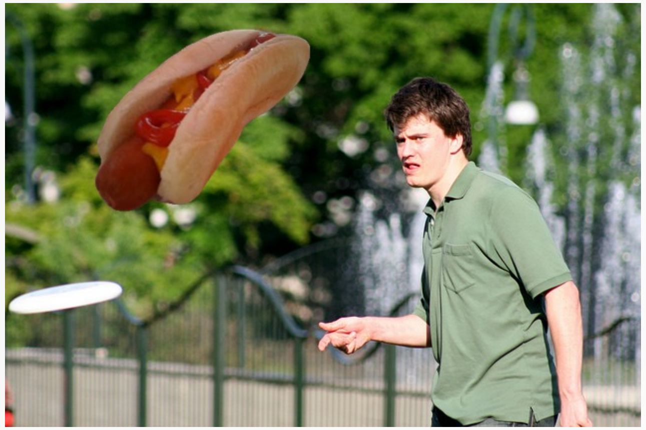


Motivation

The current automated IC testing tools always generate unnatural test cases



Original image in COCO Dataset



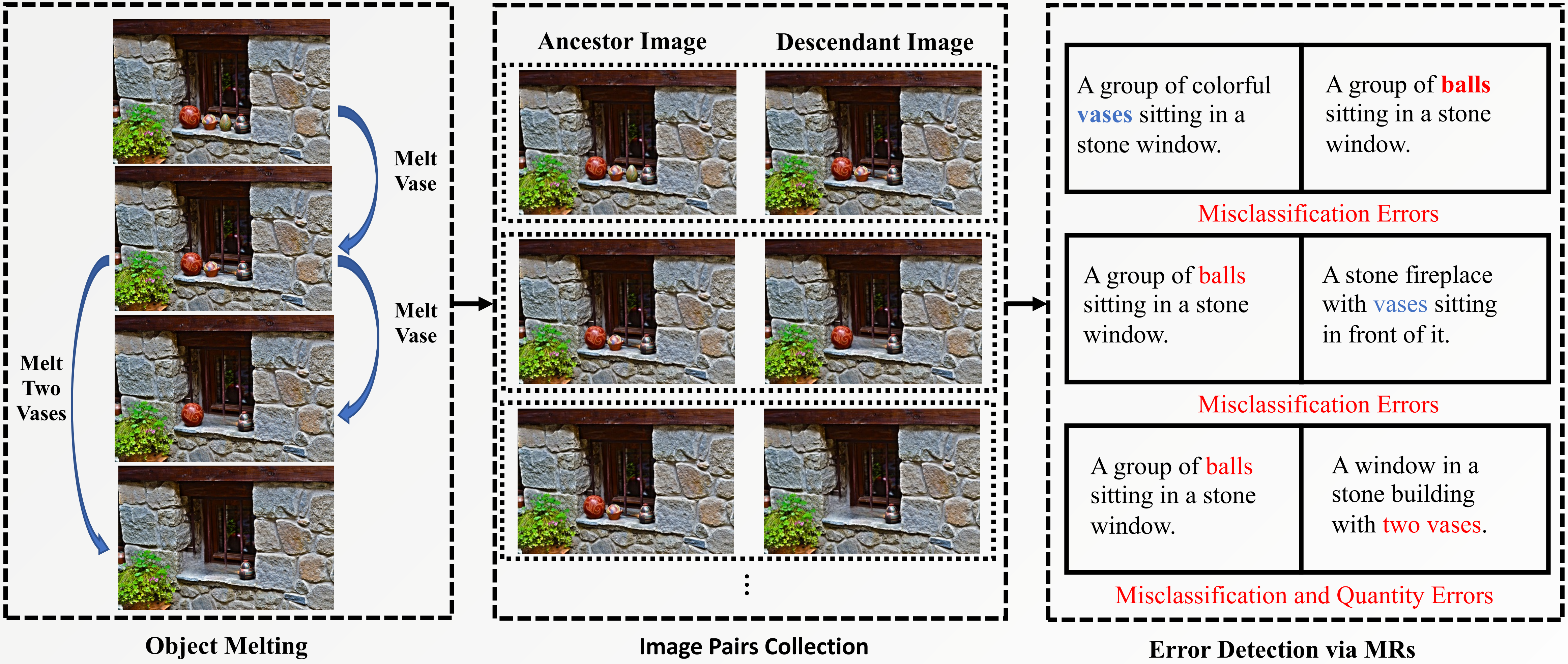
MetaIC

[1] MetaIC: Automated testing of image captioning systems

Motivation

Many critical real-world scenarios rely on captioning ability for natural images (e.g., assisting visually impaired people)

Overview of ROME



A group of colorful **vases** sitting in a stone window.	A group of **balls** sitting in a stone window.

Misclassification Errors

A group of balls sitting in a stone window.	A stone fireplace with vases sitting in front of it.

Misclassification Errors

A group of balls sitting in a stone window.	A window in a stone building with two vases.

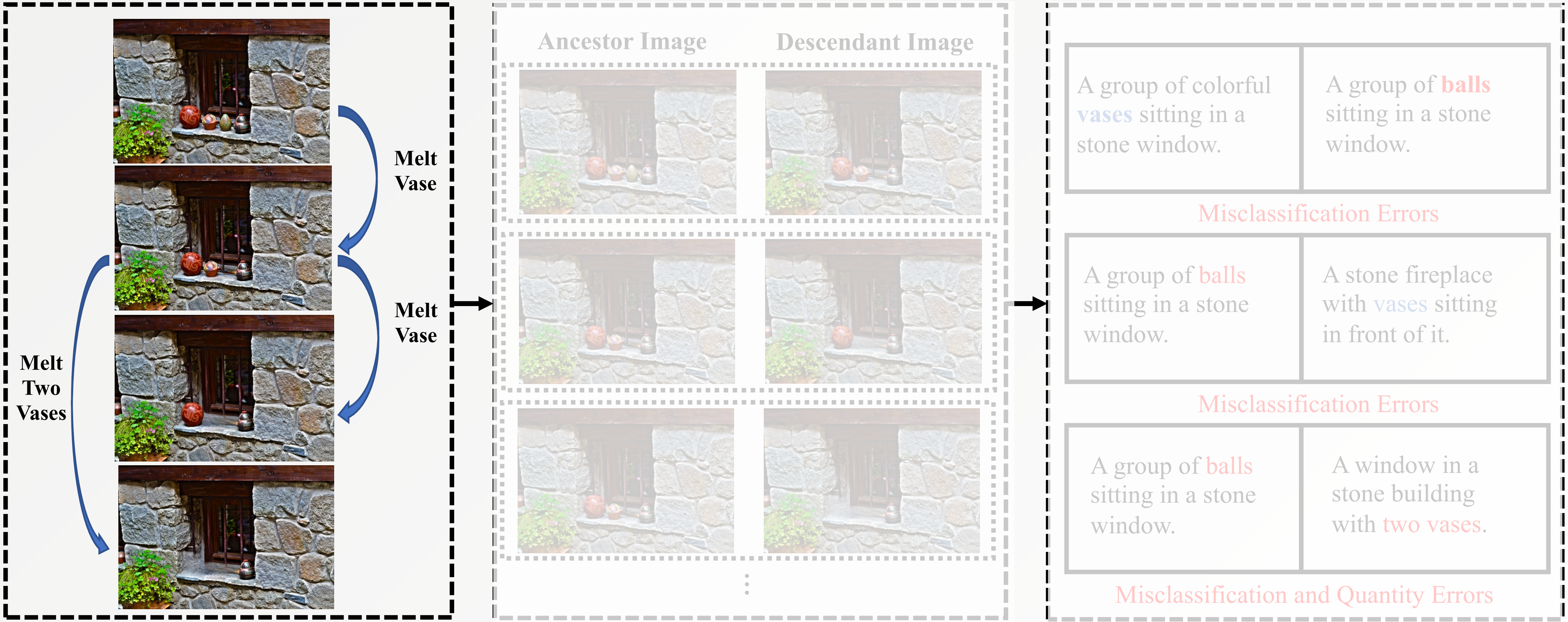
Misclassification and Quantity Errors

Object Melting

Image Pairs Collection

Error Detection via MRs

Overview of ROME



A group of colorful vases sitting in a stone window.	A group of balls sitting in a stone window.

Misclassification Errors

A group of balls sitting in a stone window.	A stone fireplace with vases sitting in front of it.

Misclassification Errors

A group of balls sitting in a stone window.	A window in a stone building with two vases.

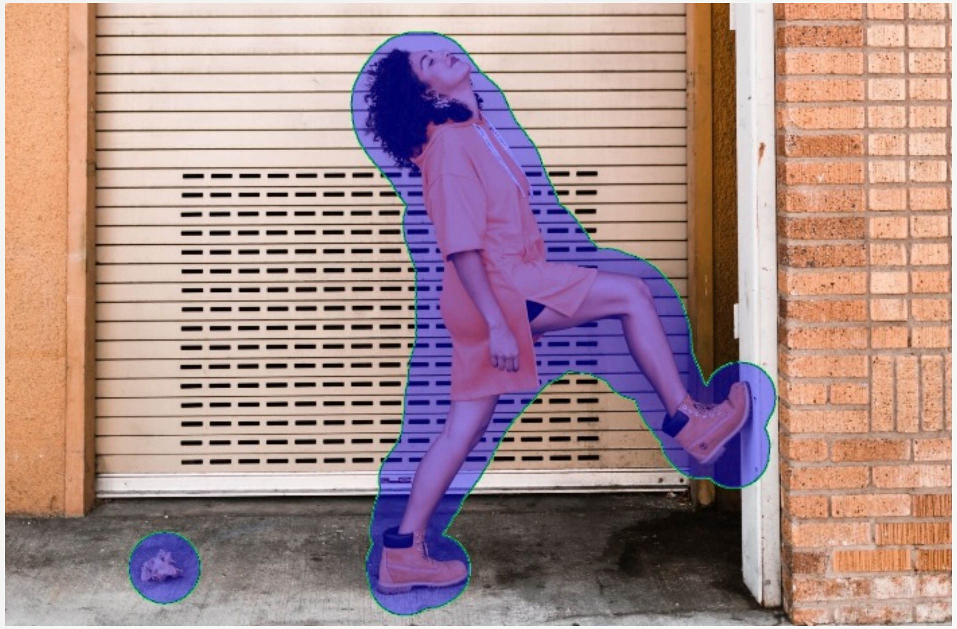
Misclassification and Quantity Errors

Object Melting

Image Pairs Collection

Error Detection via MRs

Object Melting with LaMa [2]



Original image

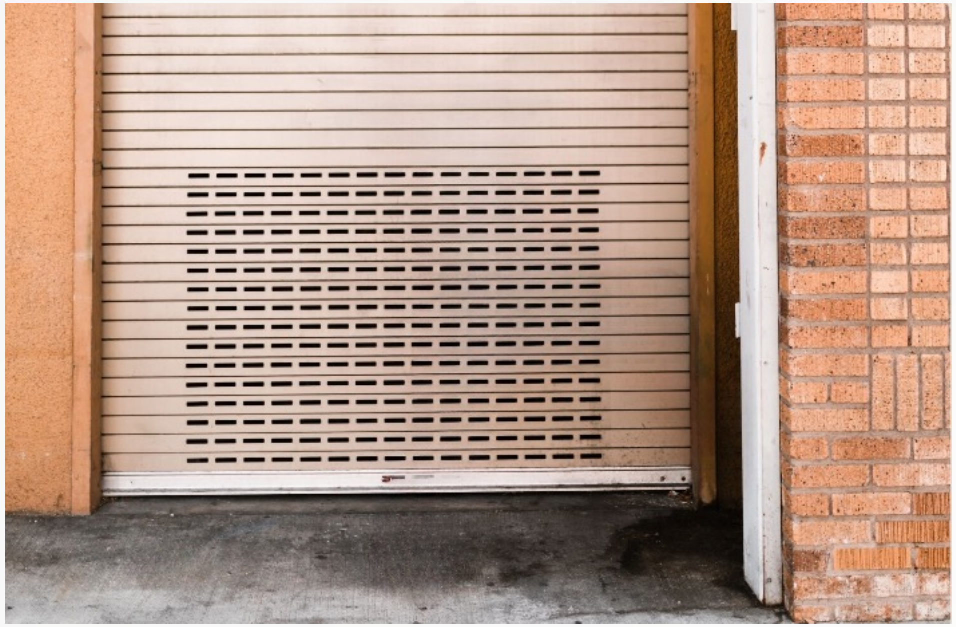
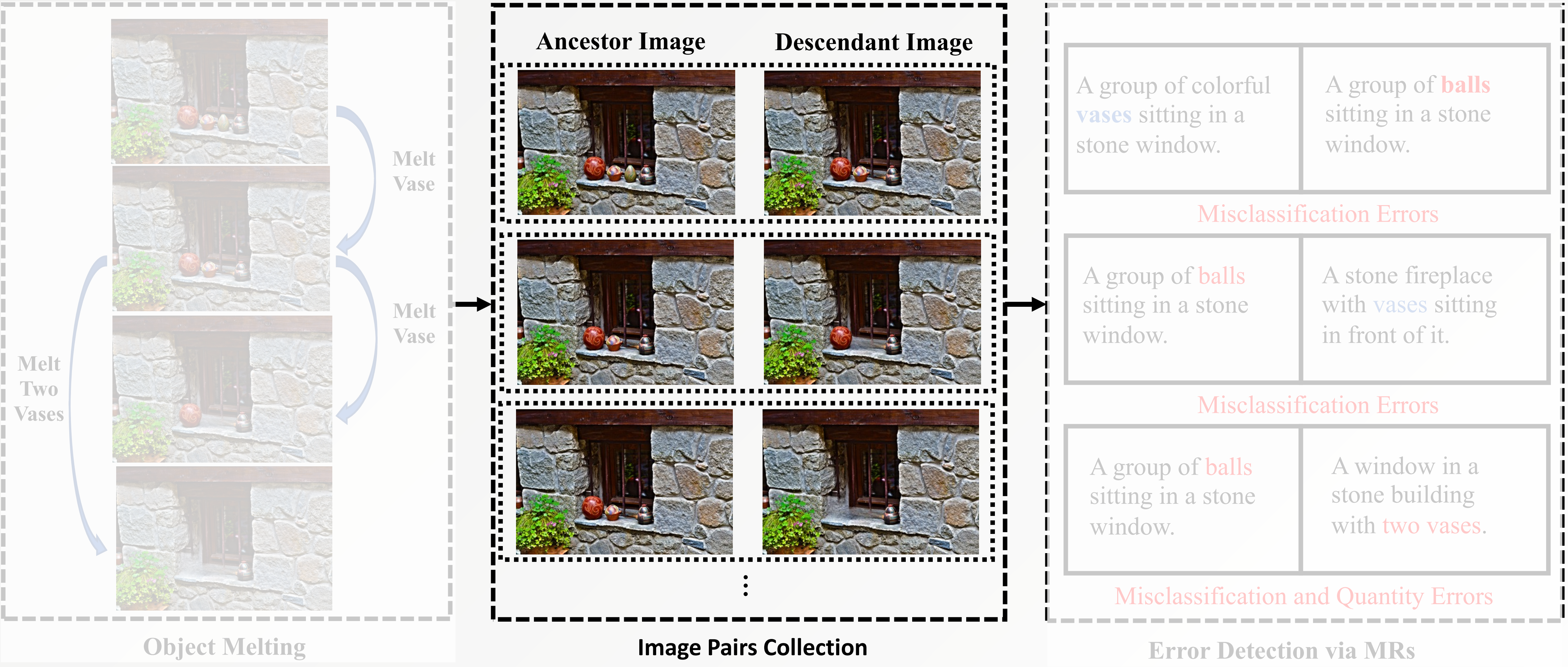


Image with the objects removed

[2] LaMa: Resolution-robust Large Mask Inpainting with Fourier Convolutions

Overview of ROME



Object Melting

Image Pairs Collection

Error Detection via MRs

A group of colorful vases sitting in a stone window.	A group of balls sitting in a stone window.

Misclassification Errors

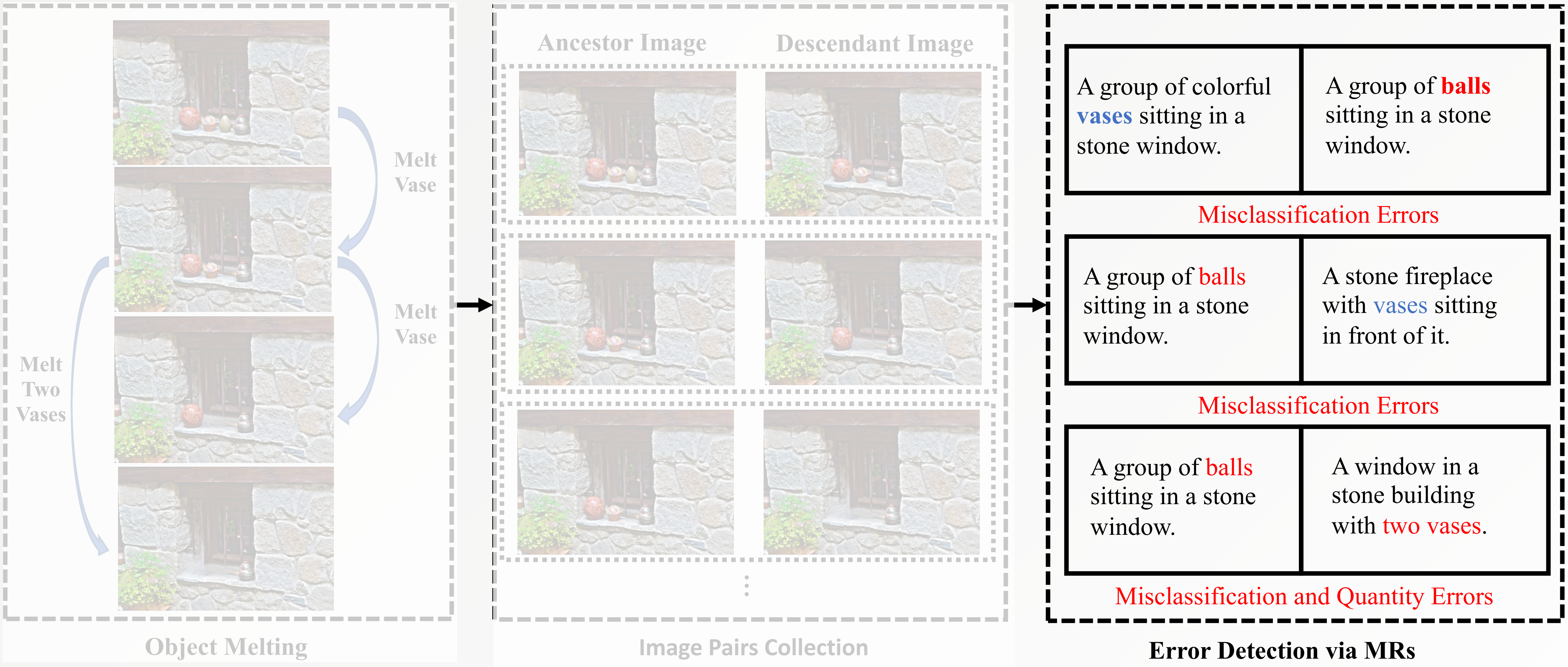
A group of balls sitting in a stone window.	A stone fireplace with vases sitting in front of it.

Misclassification Errors

A group of balls sitting in a stone window.	A window in a stone building with two vases.

Misclassification and Quantity Errors

Overview of ROME



A group of colorful **vases** sitting in a stone window.

A group of **balls** sitting in a stone window.

Misclassification Errors

A group of balls sitting in a stone window.

A stone fireplace with vases sitting in front of it.

Misclassification Errors

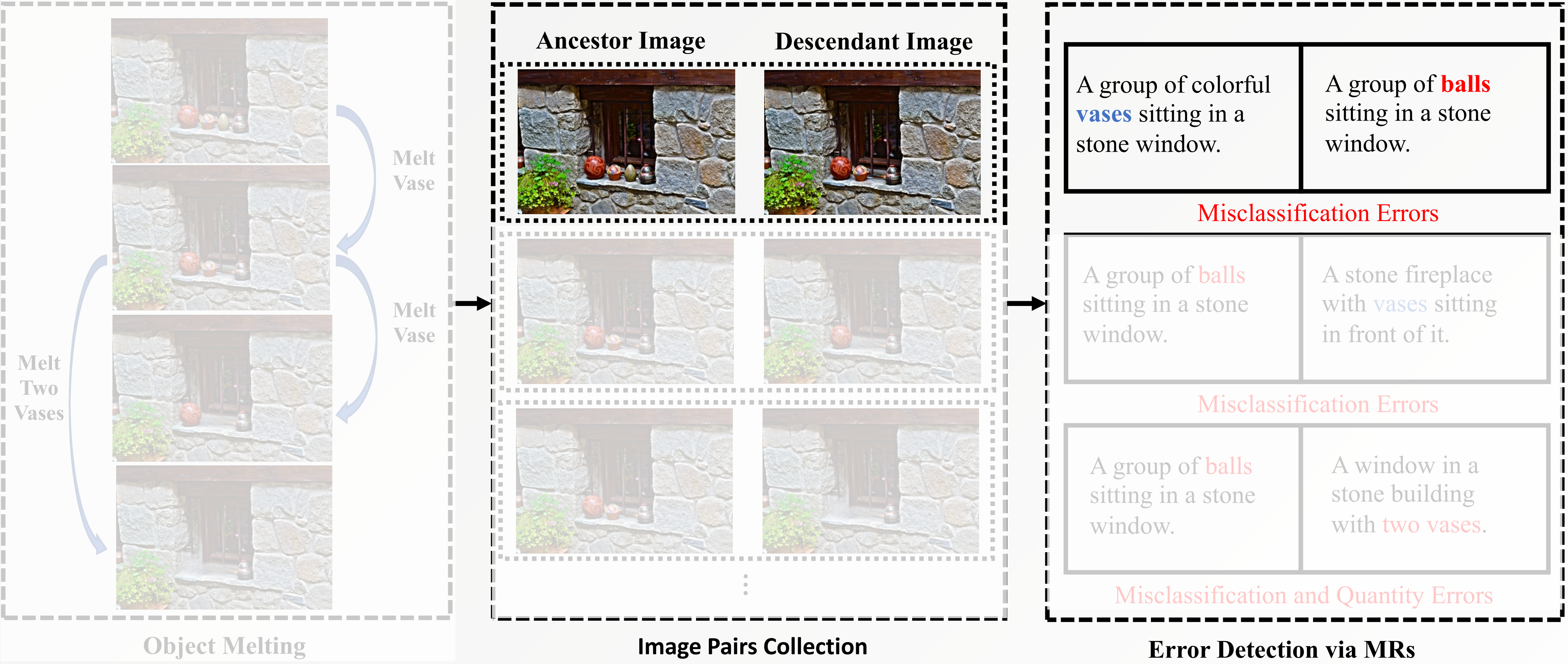
A group of balls sitting in a stone window.

A window in a stone building with two vases.

Misclassification and Quantity Errors

Error Detection via MRs

Overview of ROME



Object Melting

Image Pairs Collection

Error Detection via MRs

A group of colorful **vases** sitting in a stone window.	A group of **balls** sitting in a stone window.

Misclassification Errors

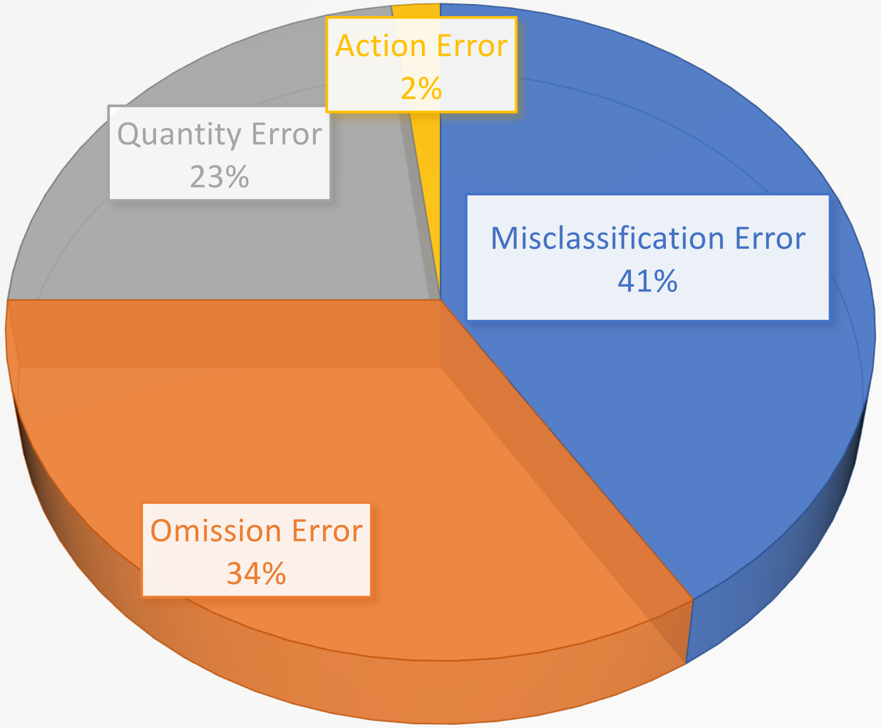
A group of balls sitting in a stone window.	A stone fireplace with vases sitting in front of it.

Misclassification Errors

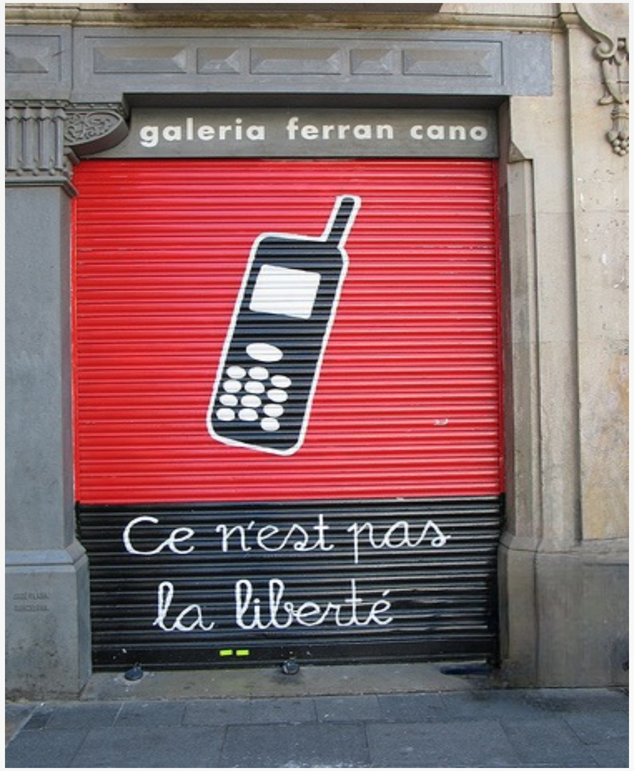
A group of balls sitting in a stone window.	A window in a stone building with two vases.

Misclassification and Quantity Errors

Categories of Captioning Errors

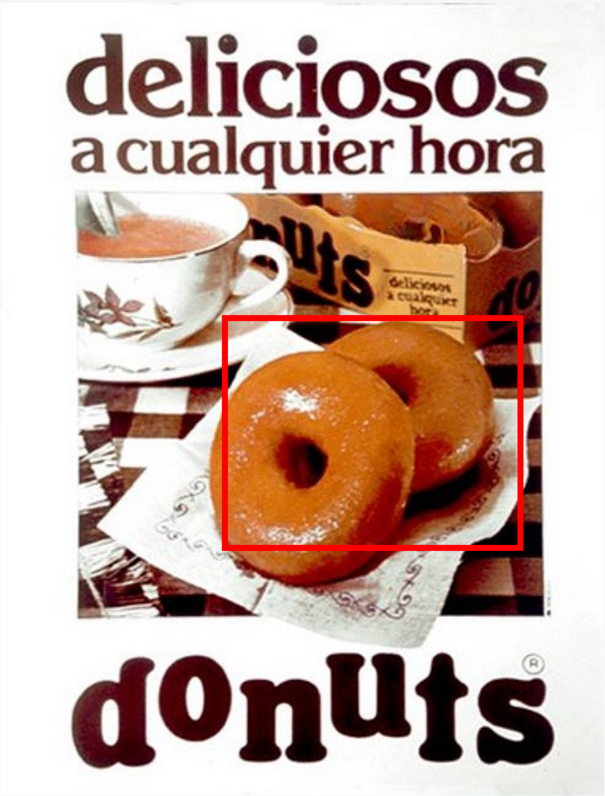


Misclassification Error



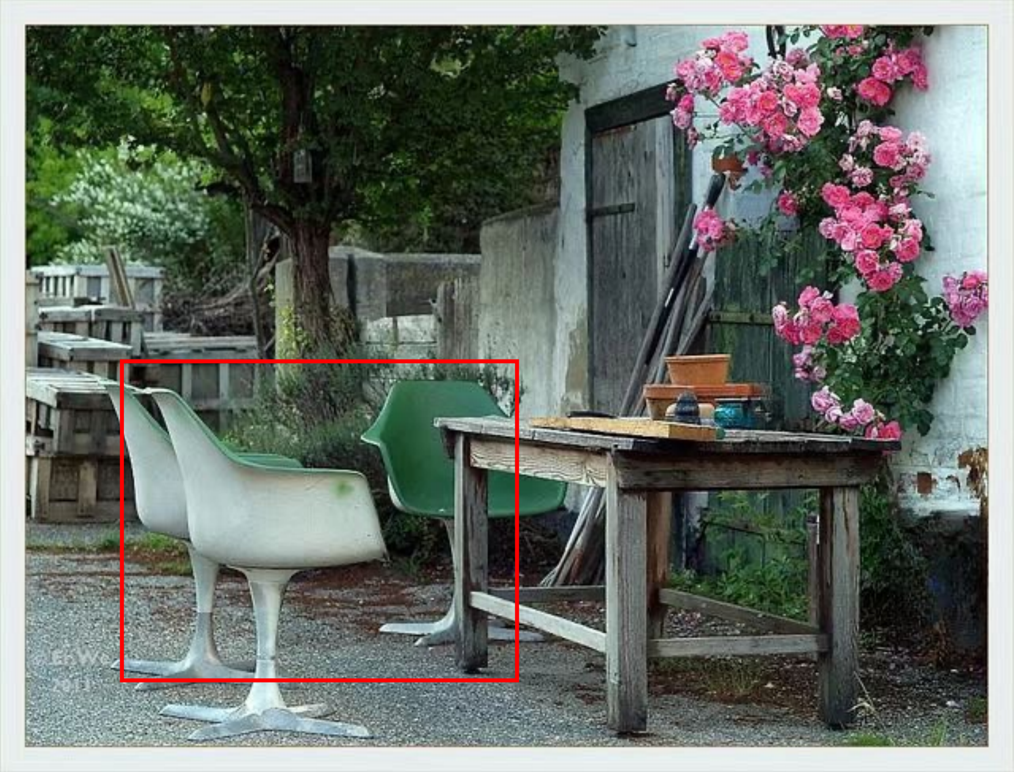
a red door with a **refrigerator** on the side of it.

Quantity Error



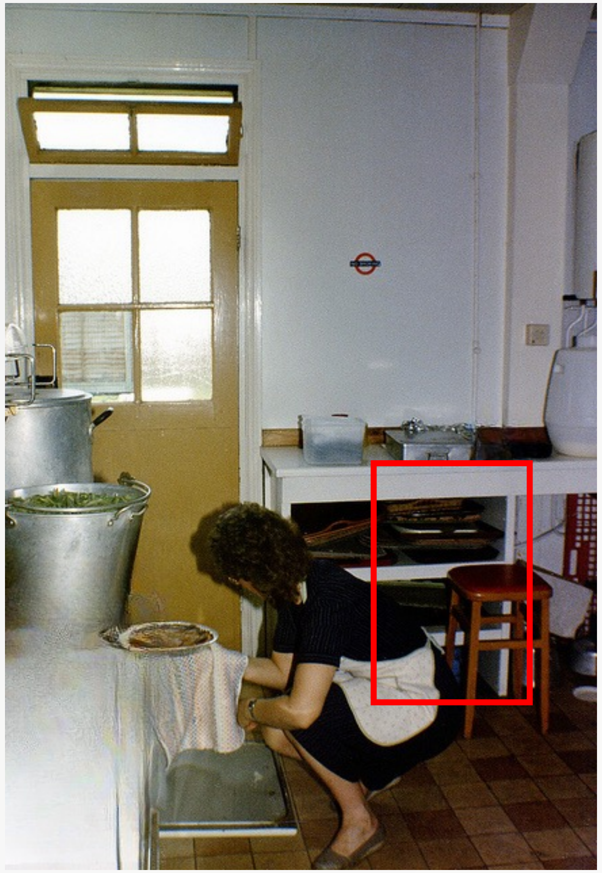
a picture of a donut and a cup of coffee.

Omission Error



a table with a vase of flowers on it.

Action Error



a person sitting on a chair in a room.

User Study on the Naturalness of Images

Evaluation Criteria:

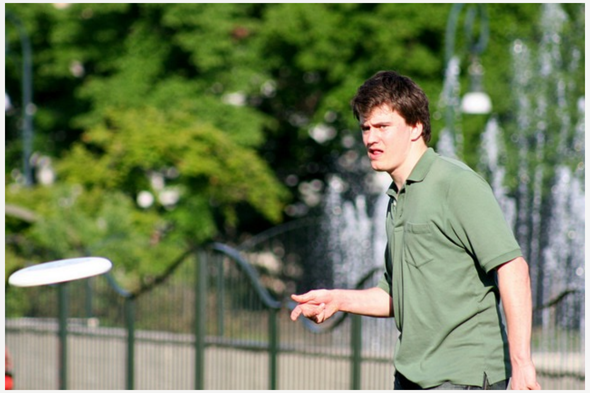
“4” denotes that the image appears to have been captured in a natural setting and appears to be a true-to-life representation of nature.

“3” denotes that the image may not be entirely natural, but it could still have been captured in nature.

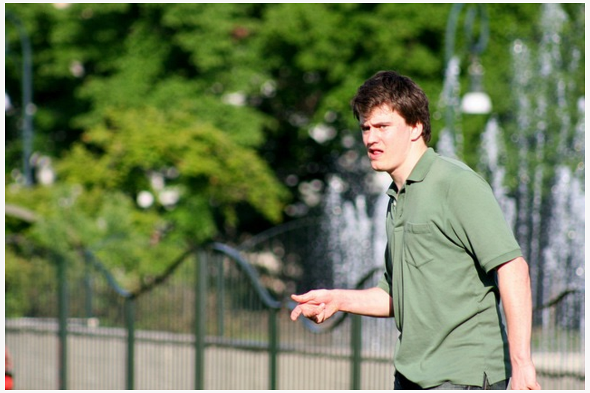
“2” denotes that the image is somewhat unnatural and would be difficult to capture in nature.

“1” denotes that the image appears to be highly unnatural and cannot be considered a representation of nature.

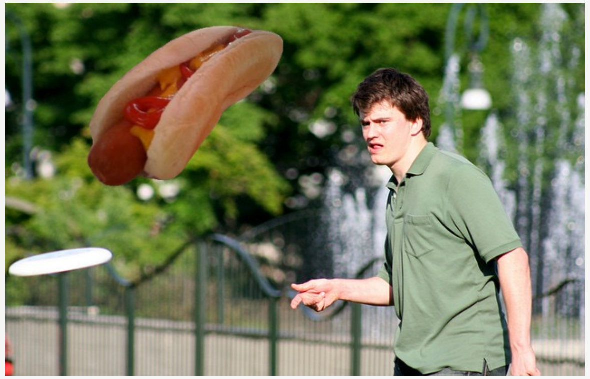
Naturalness of the Generated Image



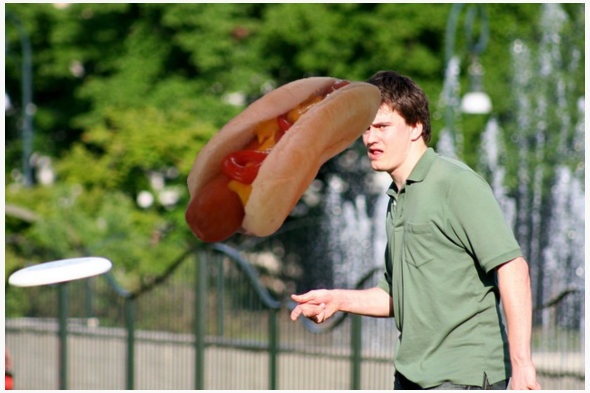
COCO Dataset



ROME

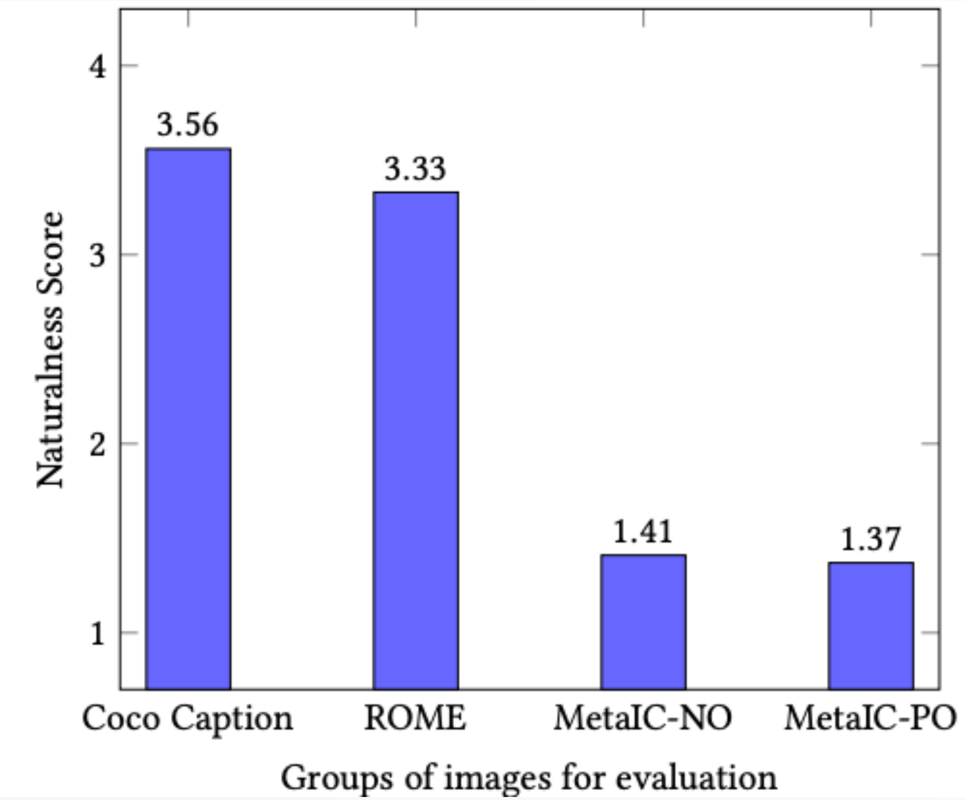


MetaIC (0%)



MetaIC (30%)

Naturalness Score by Crowd-sourcing



- MetaIC-NO: no inserted object overlapping with original ones
- MetaIC-NO: no inserted object overlapping with original ones

Precision

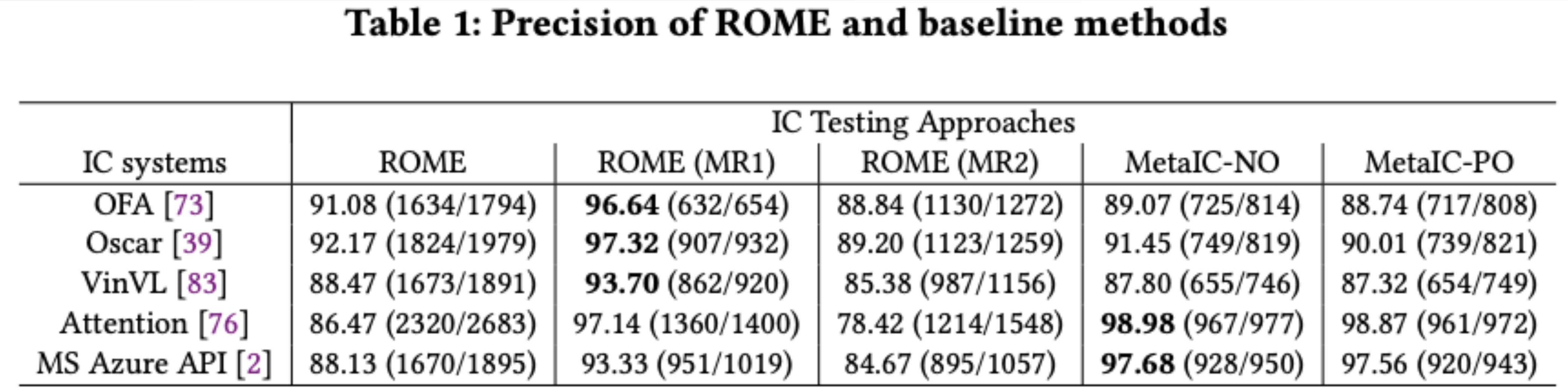
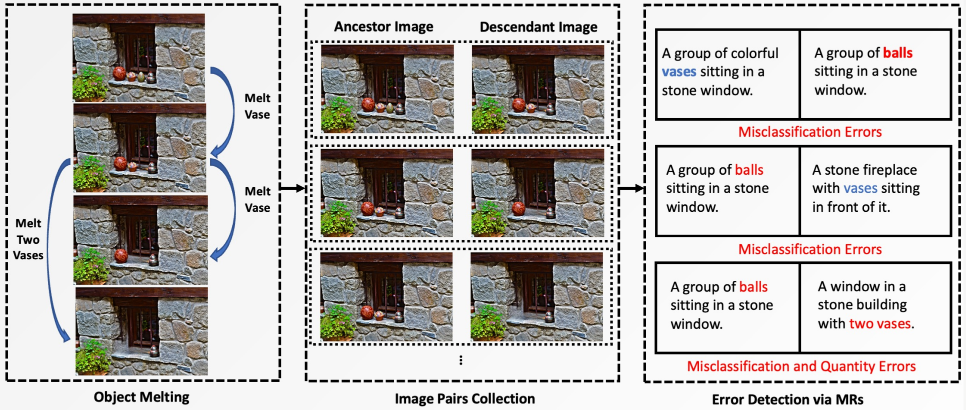


Table 1: Precision of ROME and baseline methods

	IC Testing Approaches				
IC systems	ROME	ROME (MR1)	ROME (MR2)	MetaIC-NO	MetaIC-PO
OFA [73]	91.08 (1634/1794)	**96.64** (632/654)	88.84 (1130/1272)	89.07 (725/814)	88.74 (717/808)
Oscar [39]	92.17 (1824/1979)	**97.32** (907/932)	89.20 (1123/1259)	91.45 (749/819)	90.01 (739/821)
VinVL [83]	88.47 (1673/1891)	**93.70** (862/920)	85.38 (987/1156)	87.80 (655/746)	87.32 (654/749)
Attention [76]	86.47 (2320/2683)	97.14 (1360/1400)	78.42 (1214/1548)	**98.98** (967/977)	98.87 (961/972)
MS Azure API [2]	88.13 (1670/1895)	93.33 (951/1019)	84.67 (895/1057)	**97.68** (928/950)	97.56 (920/943)

Conclusion

1 Overview of ROME



A group of colorful vases sitting in a stone window. | A group of balls sitting in a stone window.

Misclassification Errors

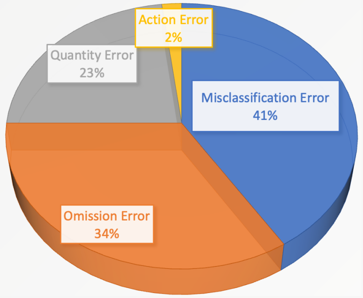
A group of balls sitting in a stone window. | A stone fireplace with vases sitting in front of it.

Misclassification Errors

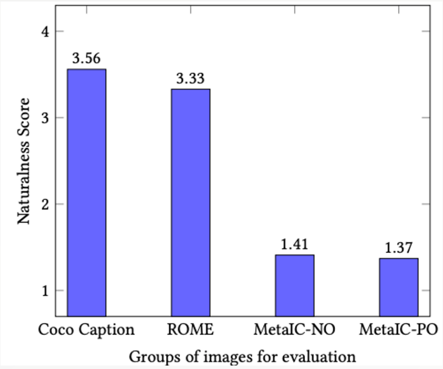
A group of balls sitting in a stone window. | A window in a stone building with two vases.

Misclassification and Quantity Errors

2 Categories of Captioning Errors



3 Naturalness Score by Crowd-sourcing



4 Precision

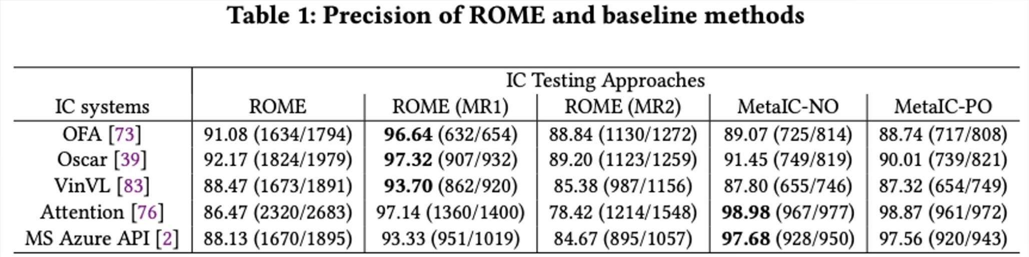


Table 1: Precision of ROME and baseline methods

IC systems	IC Testing Approaches ROME	ROME (MR1)	ROME (MR2)	MetaIC-NO	MetaIC-PO
OFA [73]	91.08 (1634/1794)	**96.64** (632/654)	88.84 (1130/1272)	89.07 (725/814)	88.74 (717/808)
Oscar [39]	92.17 (1824/1979)	**97.32** (907/932)	89.20 (1123/1259)	91.45 (749/819)	90.01 (739/821)
VinVL [83]	88.47 (1673/1891)	**93.70** (862/920)	85.38 (987/1156)	87.80 (655/746)	87.32 (654/749)
Attention [76]	86.47 (2320/2683)	97.14 (1360/1400)	78.42 (1214/1548)	**98.98** (967/977)	98.87 (961/972)
MS Azure API [2]	88.13 (1670/1895)	93.33 (951/1019)	84.67 (895/1057)	**97.68** (928/950)	97.56 (920/943)