

| Model                         | Text Retrieval  |             |             |              |             |             |              |              |             |                                    |              |              |                                   |              |              |
|-------------------------------|-----------------|-------------|-------------|--------------|-------------|-------------|--------------|--------------|-------------|------------------------------------|--------------|--------------|-----------------------------------|--------------|--------------|
|                               | Flicker30K-Full |             |             | MS-COCO-Full |             |             | MS-COCO-1K   |              |             | JourneyBench-1K<br>w/o distractors |              |              | JourneyBench-1K<br>w/ distractors |              |              |
|                               | R@1             | R@5         | R@10        | R@1          | R@5         | R@10        | R@1          | R@5          | R@10        | R@1                                | R@5          | R@10         | R@1                               | R@5          | R@10         |
| ALBEF-210M                    | 88.5            | 98.5        | 99.2        | 73.96        | 91.8        | 96.0        | 89.1         | 98.3         | 99.6        | 72.3                               | 86.1         | 91.78        | 65.36                             | 83.75        | 89.13        |
| BEiT3-674M                    | 89.5            | 98.8        | 99.4        | 64           | 86.6        | 92.2        | 81.1         | 96.6         | 98.8        | 74.1                               | 87.80        | 92.70        | 65.9                              | 86.1         | 90.9         |
| BLIP2-12B                     | 92.8            | 99.9        | 99.9        | 80.1         | 94.8        | 97.9        | 91.3         | 99.1         | 99.6        | <b>81.29</b>                       | <b>95.17</b> | <b>97.28</b> | 63.78                             | 87.76        | 92.46        |
| CLIP-430M                     | 85.3            | 97.9        | 99.1        | 58.4         | 81.5        | 88.1        | 75.6         | 93.2         | 97.5        | 70.6                               | 85.7         | 91           | 60.8                              | 83.3         | 88.5         |
| X <sup>2</sup> VLM-Large-590M | <b>98.8</b>     | <b>100</b>  | <b>100</b>  | <b>84.4</b>  | <b>96.5</b> | <b>98.5</b> | <b>93.6</b>  | <b>99.5</b>  | <b>99.9</b> | 78.54                              | 92.78        | 96.15        | 64.97                             | <b>90.47</b> | <b>94.8</b>  |
| InternVL-C-13B                | 94.7            | 99.6        | 99.9        | 74.9         | 91.3        | 95.2        | 85.34        | 96.86        | 98.84       | 78.22                              | 89.21        | 93.61        | <b>67.73</b>                      | 86.41        | 91.91        |
| InternVL-G-14B                | 95.7            | 99.7        | 99.9        | 74.9         | 91.3        | 95.2        | 87.58        | 97.64        | 99.28       | 78.52                              | 89.81        | 94.21        | 67.53                             | 86.51        | 92.61        |
| OpenCLIP-CoCa-13B             | 92.5            | 99.5        | 99.9        | 66.3         | 86.2        | 91.8        | 75.89        | 93.63        | 97.15       | 70.43                              | 85.41        | 89.61        | 60.04                             | 83.32        | 87.91        |
| Pick-a-pic-986M               | 68.5            | 89.9        | 94.9        | 44.5         | 70.2        | 80.5        | 64.5         | 89.3         | 94.4        | 62.2                               | 82.6         | 88.9         | 55.0                              | 79.7         | 86.2         |
| Model                         | Image Retrieval |             |             |              |             |             |              |              |             |                                    |              |              |                                   |              |              |
|                               | Flicker30K-Full |             |             | MS-COCO-Full |             |             | MS-COCO-1K   |              |             | JourneyBench-1K<br>w/o distractors |              |              | JourneyBench-1K<br>w/ distractors |              |              |
|                               | R@1             | R@5         | R@10        | R@1          | R@5         | R@10        | R@1          | R@5          | R@10        | R@1                                | R@5          | R@10         | R@1                               | R@5          | R@10         |
| ALBEF-210M                    | 75.9            | 92.6        | 96          | 54           | 78.99       | 87.18       | 72.28        | 94.18        | 97.54       | 66.12                              | 88.65        | 92.15        | 50.02                             | 75.46        | 82.56        |
| BEiT3-674M                    | 75.94           | 93.34       | 96.66       | 48.9         | 73.2        | 81.8        | 66.4         | 89.5         | 95.2        | 68                                 | 90.3         | 94.1         | 56.2                              | 79.9         | 85.7         |
| BLIP2-12B                     | 89.7            | 98.1        | 98.9        | 63           | 84.2        | 90.2        | 78.78        | 94.92        | 97.74       | 75.77                              | 91.66        | 94.12        | 59.97                             | 82.48        | 87.17        |
| CLIP-430M                     | 64.9            | 87.2        | 92          | 37.8         | 62.4        | 72.2        | 54.5         | 81.8         | 91          | 66.8                               | 88.8         | 92.5         | 51.2                              | 76.5         | 83.5         |
| X <sup>2</sup> VLM-Large-590M | <b>91.8</b>     | <b>98.6</b> | <b>99.5</b> | <b>67.7</b>  | <b>87.5</b> | <b>92.5</b> | <b>83.32</b> | <b>96.86</b> | <b>98.6</b> | 75.04                              | 93.16        | 95.9         | 61.02                             | 85           | 89.69        |
| InternVL-C-13B                | 81.7            | 96          | 98.2        | 54.1         | 77.3        | 84.6        | 71.43        | 91.5         | 96.28       | 75.84                              | 93.34        | 96.31        | 62.29                             | 83.44        | 89.33        |
| InternVL-G-14B                | 85              | 97          | 98.6        | 58.6         | 81.3        | 88.0        | 75.64        | 93.77        | 97.48       | <b>76.8</b>                        | <b>93.8</b>  | <b>96.4</b>  | <b>63.71</b>                      | <b>84.84</b> | <b>90.28</b> |
| OpenCLIP-CoCa-13B             | 80.4            | 95.7        | 97.7        | 51.2         | 74.2        | 82.0        | 59.30        | 85.51        | 92.78       | 65.83                              | 86.66        | 91.41        | 48.70                             | 72.56        | 80.53        |
| Pick-a-pic-986M               | 65.4            | 87.7        | 93.2        | 37.3         | 62.8        | 73.4        | 57.7         | 85.7         | 93.6        | 45.6                               | 73.6         | 81.7         | 20.3                              | 45.8         | 56.4         |

Table 1: Zero-shot evaluation of retrieval tasks on different datasets along with our proposed Journey-Bench fine-grained cross-modal retrieval datasets. The best results are highlighted in bold.

| Categories                        | Text Retrieval  |              |           |                         |                |                |                   |                 |
|-----------------------------------|-----------------|--------------|-----------|-------------------------|----------------|----------------|-------------------|-----------------|
|                                   | ALBEF-210M      | BEiT3-674M   | CLIP-430M | X <sup>2</sup> VLM-590M | InternVL-C-13B | InternVL-G-14B | OpenCLIP-CoCa-13B | Pick-a-Pic-986M |
| Incorrect physics rules           | <b>69.57</b>    | 67.39        | 60.87     | 58.70                   | 67.53          | 68.08          | 68.22             | 47.83           |
| Incorrect biological rules        | 63.38           | <b>67.61</b> | 54.93     | 53.52                   | 63.53          | 64.05          | 64.11             | 52.11           |
| Misplacement                      | 71.60           | 69.14        | 69.14     | 61.73                   | 71.66          | <b>71.81</b>   | 70.14             | 59.26           |
| Strange animal                    | <b>73.15</b>    | 73.15        | 63.89     | 61.11                   | 70.96          | 71.84          | 74.11             | 52.78           |
| Unexpected behavior               | 76.47           | 77.65        | 72.55     | 70.98                   | 75.20          | <b>78.31</b>   | 70.84             | 69.41           |
| Unusual food                      | 68.75           | 70.83        | 72.92     | 70.83                   | 72.02          | <b>74.69</b>   | 71.92             | 68.75           |
| Strange indoor objects            | 60.00           | 58.46        | 56.92     | 58.46                   | 62.33          | 62.94          | <b>63.77</b>      | 47.69           |
| Strange scene                     | 60.00           | 56.00        | 49.60     | 48.00                   | 56.38          | 57.89          | <b>61.00</b>      | 44.4            |
| Unusual construction              | 51.52           | <b>53.03</b> | 43.94     | 45.45                   | 52.15          | 52.87          | 52.39             | 43.94           |
| Incorrect usage                   | 60.00           | <b>74.29</b> | 71.43     | 60.00                   | 70.97          | 70.49          | 72.14             | 65.71           |
| Unusual attributes or accessories | 60.70           | 60.95        | 58.46     | 58.21                   | 63.28          | <b>63.59</b>   | 55.74             | 52.24           |
| Categories                        | Image Retrieval |              |           |                         |                |                |                   |                 |
|                                   | ALBEF-210M      | BEiT3-674M   | CLIP-430M | X <sup>2</sup> VLM-590M | InternVL-C-13B | InternVL-G-14B | OpenCLIP-CoCa-13B | Pick-a-Pic-986M |
| incorrect physics rules           | 48.70           | 52.17        | 53.91     | 57.39                   | 59.70          | <b>60.52</b>   | 53.20             | 27.83           |
| incorrect biological rules        | 46.20           | 55.21        | 50.70     | 59.72                   | 59.78          | <b>60.48</b>   | 53.02             | 14.93           |
| misplacement                      | 56.79           | 65.43        | 57.78     | 59.26                   | 66.72          | <b>67.44</b>   | 53.18             | 20.99           |
| strange animal                    | 55.56           | 60.74        | 47.04     | 66.11                   | 63.60          | <b>64.21</b>   | 60.47             | 20.00           |
| unexpected behavior               | 63.37           | 67.37        | 64.16     | 68.63                   | 72.28          | <b>72.81</b>   | 61.14             | 22.59           |
| unusual food                      | 60.83           | 73.33        | 70.83     | 74.17                   | 76.23          | <b>76.97</b>   | 59.88             | 24.58           |
| strange indoor objects            | 49.54           | 54.77        | 47.38     | 52.92                   | 57.50          | <b>58.00</b>   | 49.61             | 12.62           |
| strange scene                     | 42.64           | 47.52        | 40.08     | 47.52                   | 50.76          | <b>51.81</b>   | 43.26             | 14.48           |
| unusual construction              | 27.58           | 38.18        | 36.97     | 36.67                   | 41.55          | <b>42.36</b>   | 28.08             | 20.00           |
| incorrect usage                   | 64.57           | 74.29        | 71.43     | 68.57                   | 76.50          | <b>77.12</b>   | 62.84             | 24.00           |
| unusual attributes or accessories | 48.51           | 52.89        | 49.65     | 53.98                   | 57.78          | <b>59.28</b>   | 49.13             | 19.75           |

Table 2: Zero-shot detailed results (R@1) of Retrieval across categories on our proposed Journey-Bench dataset.