

000 001 002 003 004 005 006 007 LEARNING TO GENERATE STYLIZED HANDWRITTEN 008 TEXT VIA A UNIFIED REPRESENTATION OF STYLE, 009 CONTENT, AND NOISE 010 011

012 **Anonymous authors**
013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029
030
031
032
033
034
035
036
037
038
039
040
041
042
043
044
045
046
047
048
049
050
051
052
053
054
055
056
057
058
059
060
061
062
063
064
065
066
067
068
069
070
071
072
073
074
075
076
077
078
079
080
081
082
083
084
085
086
087
088
089
090
091
092
093
094
095
096
097
098
099
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
559
560
561
562
563
564
565
566
567
568
569
569
570
571
572
573
574
575
576
577
578
579
579
580
581
582
583
584
585
586
587
588
589
589
590
591
592
593
594
595
596
597
598
599
599
600
601
602
603
604
605
606
607
608
609
609
610
611
612
613
614
615
616
617
618
619
619
620
621
622
623
624
625
626
627
628
629
629
630
631
632
633
634
635
636
637
638
639
639
640
641
642
643
644
645
646
647
648
649
649
650
651
652
653
654
655
656
657
658
659
659
660
661
662
663
664
665
666
667
668
669
669
670
671
672
673
674
675
676
677
678
679
679
680
681
682
683
684
685
686
687
688
689
689
690
691
692
693
694
695
696
697
698
699
699
700
701
702
703
704
705
706
707
708
709
709
710
711
712
713
714
715
716
717
718
719
719
720
721
722
723
724
725
726
727
728
729
729
730
731
732
733
734
735
736
737
738
739
739
740
741
742
743
744
745
746
747
748
749
749
750
751
752
753
754
755
756
757
758
759
759
760
761
762
763
764
765
766
767
768
769
769
770
771
772
773
774
775
776
777
778
779
779
780
781
782
783
784
785
786
787
788
789
789
790
791
792
793
794
795
796
797
798
799
799
800
801
802
803
804
805
806
807
808
809
809
810
811
812
813
814
815
816
817
818
819
819
820
821
822
823
824
825
826
827
828
829
829
830
831
832
833
834
835
836
837
838
839
839
840
841
842
843
844
845
846
847
848
849
849
850
851
852
853
854
855
856
857
858
859
859
860
861
862
863
864
865
866
867
868
869
869
870
871
872
873
874
875
876
877
878
879
879
880
881
882
883
884
885
886
887
888
889
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
909
910
911
912
913
914
915
916
917
918
919
919
920
921
922
923
924
925
926
927
928
929
929
930
931
932
933
934
935
936
937
938
939
939
940
941
942
943
944
945
946
947
948
949
949
950
951
952
953
954
955
956
957
958
959
959
960
961
962
963
964
965
966
967
968
969
969
970
971
972
973
974
975
976
977
978
979
979
980
981
982
983
984
985
986
987
988
989
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1909
1910
1911
1912
1913
1914
1915
1916
1917<br

054 plications in assistive technology, personalized rendering, font design, historical manuscript restora-
 055 tion, and writer identification.

056 Diffusion models have recently emerged as the dominant paradigm for offline handwritten text syn-
 057 thesis, surpassing GAN-based approaches in generation quality. Early attempts at style condition-
 058 ing (e.g., WordStylist (Nikolaïdou et al., 2023), GC-DDPM (Ding et al., 2023), CTIG-DM (Zhu
 059 et al., 2023)) relied on fixed writer IDs, which constrained stylistic diversity. Subsequent methods
 060 introduced dedicated style encoders trained with tailored objectives to capture broader stylistic vari-
 061 ations. For instance, One-DM (Dai et al., 2024) employs a Laplacian contrastive loss to emphasize
 062 fine-grained features, while DiffusionPen (Nikolaïdou et al., 2024) integrates triplet and classifica-
 063 tion losses to enhance style discrimination. With respect to content fidelity, TGC-Diff (Wang et al.,
 064 2025) proposes a high-frequency mask loss to preserve structural details. While these approaches
 065 improve model performance, they still handle style, content, and noise as separate factors, each
 066 constrained by manually crafted losses, which increases the difficulty of optimization.

067 To efficiently integrate diffusion models into HTG systems, it is crucial to examine the role of aux-
 068 iliary style and content encoders. As illustrated in Figure 1(a), One-DM fuses style and content
 069 features to provide useful guidance for style learning but fails to capture fine-grained structures
 070 and spatial information. In contrast, TGC-Diff in Figure 1(b) constructs content features within the
 071 same latent space as noise, facilitating a seamless connection between content and noise repres-
 072 entations and enhancing structural fidelity and spatial consistency. If content and noise can be jointly
 073 represented in a shared latent space, is it possible to design a single unified diffusion model that
 074 simultaneously processes style, content, and noise? Such a framework would not only obviate the
 075 need for redundant encoders with complex handcrafted losses, but also improve performance by
 076 enabling efficient interaction within a common latent representation.

077 Building on this insight, we introduce InkSpire, a diffusion transformer model for stylized handwrit-
 078 ing generation, which is “inspired” by in-context latent “ink” tokens. As illustrated in Figure 1(c),
 079 InkSpire leverages a shared latent space for style, content, and noise, replacing separate encoders
 080 with a streamlined architecture that facilitates effective feature interaction. Moreover, recent ad-
 081 vances in large text-to-image diffusion models have demonstrated strong in-context generation ca-
 082 pabilities. To apply this unified modeling ability to our HTG task, we design a multi-line masked
 083 infilling strategy for training and remove the text encoder to enable purely visual conditioning. In
 084 addition, we revise the positional encoding mechanism to support multi-line generation of arbitrary
 085 length and train the model on a mixed Chinese–English dataset, thereby enabling bilingual stylized
 086 handwriting synthesis. By integrating these innovations, InkSpire achieves high-fidelity and stylis-
 087 tically diverse handwriting generation and editing, while streamlining the overall training process.

088 In summary, the main contributions of this paper are as follows:

- 090 • We propose InkSpire, a novel handwriting generation framework that unifies the model-
 091 ing of style, content, and noise without relying on explicit style or content encoders. By
 092 leveraging in-context generation with diffusion transformer models, InkSpire simplifies the
 093 training pipeline while preserving high fidelity and stylistic diversity.
- 094
- 095 • We introduce a multi-line masked infilling strategy that enables the model to be trained
 096 directly on raw multi-line text images, without requiring complex data preprocessing. To-
 097 gether with a revised positional encoding scheme, InkSpire supports the generation of
 098 multi-line handwriting of arbitrary length as well as fine-grained, character-level editing.
- 099
- 100 • We enable bilingual handwriting generation within a single model by training InkSpire on
 101 a mixed Chinese–English dataset. This design allows the synthesis of high-quality hand-
 102 written text in both English and Chinese scripts, thereby overcoming the constraints of
 103 language-specific handwriting systems.
- 104
- 105 • We conduct extensive experiments on the ICDAR2013 and IAM datasets, demonstrating
 106 that InkSpire produces handwriting with superior structural accuracy and stylistic consis-
 107 tency, outperforming other methods in both qualitative and quantitative evaluations.

108

2 RELATED WORK

110 Handwritten data is commonly categorized into two modalities: online trajectory sequences and
 111 offline static images. Online handwriting, which captures the dynamic pen trajectory during the
 112 writing process, has been widely studied with various generative models, including RNN-based
 113 approaches (Kotani et al., 2020; Zhao et al., 2020; Zhang et al., 2017), Transformer-based architec-
 114 tures (Dai et al., 2023), and diffusion-based methods (Luhman & Luhman, 2020; Ren et al., 2023).
 115 In contrast, offline handwriting represents static visual appearances, conveying natural characteris-
 116 tics such as stroke thickness, curvature, and ink density.

117

2.1 OFFLINE HANDWRITING GENERATION

118 Early approaches to offline handwriting synthesis predominantly relied on Generative Adversarial
 119 Networks (Alonso et al., 2019; Xie et al., 2021; Gan & Wang, 2021; Kong et al., 2022; Liu et al.,
 120 2022), where adversarial training was used to generate visually plausible text. Later, transformer-
 121 based models such as HWT (Bhunia et al., 2021) and VATr (Pippi et al., 2023a) introduced hybrid
 122 CNN-transformer designs that enhanced style representation learning and improved generalization.

123 More recently, diffusion models have emerged as the dominant paradigm, synthesizing handwriting
 124 through iterative denoising with fine-grained control over style and content. State-of-the-art meth-
 125 ods (Wang et al., 2025; Pippi et al., 2025; Dai et al., 2024; Nikolaidou et al., 2023; Ding et al., 2023;
 126 Zhu et al., 2023) demonstrate strong style adaptation from only a few references, enabling diverse
 127 and personalized handwriting generation with minimal supervision.

128

2.2 IN-CONTEXT GENERATION

129 Recent advances in diffusion-based generative modeling increasingly adopt in-context generation
 130 for controllable and personalized image synthesis. Early works such as InstructPix2Pix (Brooks
 131 et al., 2023) and its successors (Boesel & Rombach, 2024) fine-tuned diffusion models with syn-
 132 thetic instruction-response pairs for diverse editing tasks, marking initial attempts to align genera-
 133 tion with user intent. Subsequently, instruction-driven editors including Emu Edit (Sheynin et al.,
 134 2024), OmniGen (Xiao et al., 2025), HiDream-II (Cai et al., 2025), and ICEdit (Zhang et al., 2025)
 135 leveraged refined datasets and task-specific architectural enhancements to improve alignment and
 136 fidelity. Huang et al. (2024) extend this paradigm by introducing task-specific LoRA branches
 137 within diffusion transformers. Although unified models for both editing and generation have been
 138 explored in printed document understanding (Tang et al., 2023; Chen et al., 2023), our approach is
 139 the first to bring in-context modeling strength of unified editing-and-generation framework into the
 140 handwriting domain.

141

3 METHODS

142

3.1 PRELIMINARY

143

3.1.1 DATA NOTION

144 The multi-line handwritten dataset is represented as a collection of transcribed image-text pairs
 145 $\{(\mathbf{X}, \mathbf{C})\}$, where \mathbf{X} denotes a stylized text-line image and \mathbf{C} denotes the corresponding textual con-
 146 tent. Let $\mathbf{C} = (c^1, c^2, \dots, c^n)$ denote a text sequence of length n , where each element corresponds
 147 to either a character in logographic writing systems (e.g., Chinese) or a word in phonemic writing
 148 systems (e.g., English).

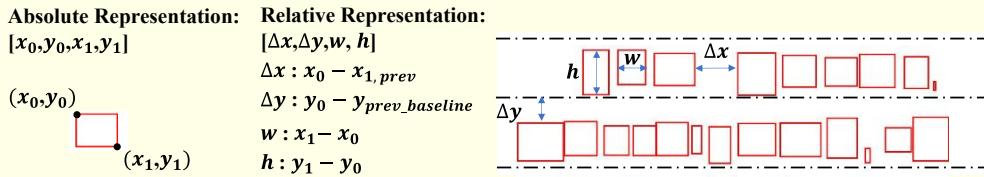
149

3.1.2 TASK ANALYSIS

150 The objective of handwritten text-line generation is to synthesize text-line images that accurately
 151 convey the given textual content \mathbf{C} while faithfully imitating the writing style of a target author,
 152 as specified by style reference samples \mathbf{X}_s . Formally, the task can be expressed as modeling the
 153 conditional distribution: $p(\mathbf{X} | \mathbf{C}, \mathbf{X}_s)$.

162

163



164

165

166

167

168

169

170

Figure 2: This figure illustrates the transformation from absolute to relative positional representations in layout information. All positions are normalized by the original paragraph image width.

171

172

173

174

However, directly modeling this distribution is challenging due to the lack of explicit character positions. Following prior work Wang et al. (2025); Yao et al. (2025), we introduce a content image \mathbf{X}_c , where each character in the text-line image \mathbf{X} is replaced with a glyph rendered in a standard font according to the stylized layout. This allows us to model the joint distribution of the text-line image and its content image: $p(\mathbf{X}, \mathbf{X}_c | \mathbf{C}, \mathbf{X}_s)$.

175

176

177

178

179

180

181

182

183

In line with Ren et al. (2023), the joint distribution can be factorized as:

184

185

186

187

$$\begin{aligned} p(\mathbf{X}, \mathbf{X}_c | \mathbf{C}, \mathbf{X}_s) &= p(\mathbf{X}_c | \mathbf{C}, \mathbf{X}_s) p(\mathbf{X} | \mathbf{C}, \mathbf{X}_s, \mathbf{X}_c) \\ &= p(\mathbf{X}_c | \mathbf{C}, \mathbf{X}_s) p(\mathbf{X} | \mathbf{X}_s, \mathbf{X}_c). \end{aligned} \quad (1)$$

Accordingly, the overall generation process decomposes into two components: (i) layout generation model $p(\mathbf{X}_c | \mathbf{C}, \mathbf{X}_s)$, and (ii) image generation model $p(\mathbf{X} | \mathbf{X}_s, \mathbf{X}_c)$.

3.2 LAYOUT GENERATION

Given an input character sequence $\mathbf{C} = (c^1, c^2, \dots, c^n)$, we define a corresponding layout sequence $\mathbf{B} = (b^1, b^2, \dots, b^n)$, where each b^i denotes the bounding box of character c^i . Specifically, b^i is parameterized by four normalized values that capture the character’s [width w , height h , horizontal offset from the previous character Δx , vertical offset from the baseline of the preceding text-line Δy], as illustrated in Figure 2. We explore three strategies for modeling layout representations:

188

189

190

191

192

193

194

- Autoregressive Modeling

195

196

197

$$p(b^i | [(b^1, c^1), (b^2, c^2), \dots, (b^{i-1}, c^{i-1})], c^i, \mathbf{X}_s). \quad (2)$$

198

199

- Masked Layout Modeling

200

201

202

203

$$p(b^i, i \in \mathcal{M} | [(b^j, c^j), j \notin \mathcal{M}], [c^k, k \in \mathcal{M}], \mathbf{X}_s). \quad (3)$$

204

205

- Masked Modeling with Conditional Flow Matching

206

207

$$p(b_{t-1}^i, i \in \mathcal{M} | [(b^j, c^j), j \notin \mathcal{M}], [(b_t^k, c^k), k \in \mathcal{M}], \mathbf{X}_s), \quad (4)$$

208

209

210

211

212

213

214

215

where \mathcal{M} denotes the index set of masked positions and t denotes the timestep. Specifically, the autoregressive model employs several transformer decoder layers, while both the masked layout model and its CFM variant utilize several transformer encoder layers. During training, the autoregressive and masked layout models are optimized by minimizing the average L1 distance between the predicted and ground-truth bounding box parameters (width, height, horizontal offset, and vertical offset) for the relevant characters. For the CFM model, the training loss is defined as the average L1 distance between the predicted and ground-truth velocity of these four layout parameters at each timestep. After training, we render the content image \mathbf{X}_c based on the predicted layout sequence \mathbf{B} and a chosen standard font, thereby successfully modeling the conditional distribution $p(\mathbf{X}_c | \mathbf{C}, \mathbf{X}_s)$.

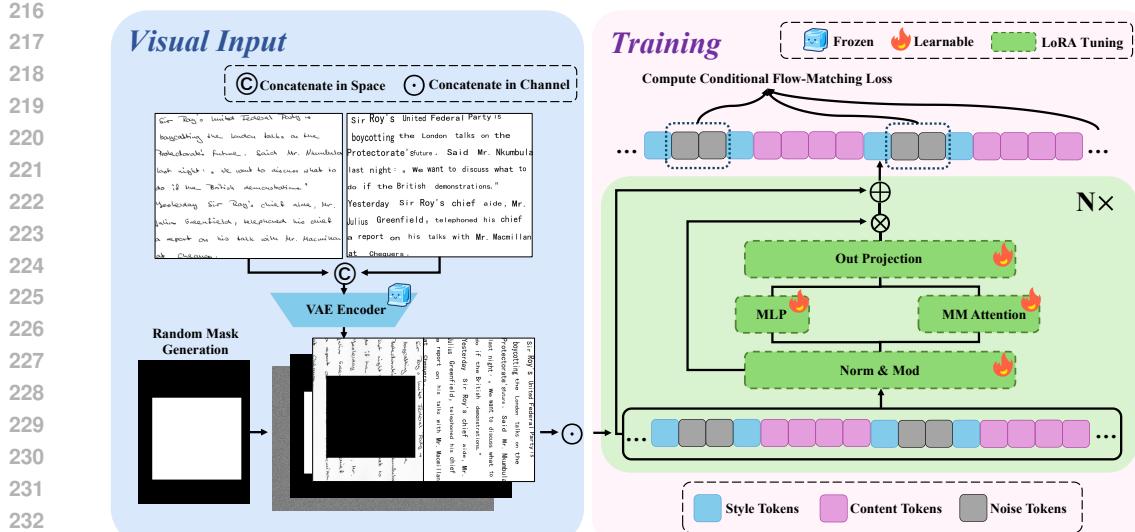


Figure 3: Overview of InkSpire. Our model achieves unified modeling by simply concatenating \mathbf{X} with \mathbf{X}_c and training directly on multi-line images with mask noise prediction. No additional encoders or extra designed loss are needed.

3.3 IMAGE GENERATION

3.3.1 MULTI-LINE MASKED INFILLING STRATEGY

Leveraging the layout generation model, the original image-text pairs $\{(\mathbf{X}, \mathbf{C})\}$ are transformed into image-content pairs $\{(\mathbf{X}, \mathbf{X}_c)\}$. Previous approaches typically construct paired target and style images $\{(\mathbf{X}_{tar}, \mathbf{X}_s)\}$, by cropping two distinct text-lines from a single author’s multi-line image \mathbf{X} . Images are then resized to a fixed height for training convenience. Under this setup, diffusion models are trained to learn $p(\mathbf{X}_{tar,t-1} \mid \mathbf{X}_{tar,t}, \mathbf{X}_s, \mathbf{X}_c)$, where \mathbf{X}_c or \mathbf{X}_s are typically processed by separate encoders, and thus do not share the same feature space as $\mathbf{X}_{tar,t}$. Such preprocessing is suboptimal as it (i) overly shrinks characters in highly slanted lines, (ii) introduces inconsistent distortions across lines with different slants, and (iii) discards inter-line style cues, thereby limiting generative capability and hindering resolution generalization.

To enable direct training on the original image pairs $\{(\mathbf{X}, \mathbf{X}_c)\}$, we first randomly crop fixed-size patches of size $P \times P$ from the original handwritten page images, where P is a hyperparameter controlling the patch size. On these patches, a random binary mask image \mathbf{M} of the same size is applied. This mask partitions the image into two complementary components: the masked region $\mathbf{X}_{mis} = \mathbf{M} \otimes \mathbf{X}$ and the observed context $\mathbf{X}_{ctx} = (1 - \mathbf{M}) \otimes \mathbf{X}$. With this construction, there is no longer a need to explicitly crop paired samples $\{(\mathbf{X}_{tar}, \mathbf{X}_s)\}$ from \mathbf{X} . In fact, \mathbf{X}_{mis} implicitly corresponds to \mathbf{X}_{tar} , while \mathbf{X}_{ctx} serves the role of \mathbf{X}_s . Consequently, the training objective is reformulated to $p(\mathbf{X}_{mis,t-1} \mid \mathbf{X}_{mis,t}, \mathbf{X}_{ctx}, \mathbf{X}_c)$, thereby eliminating the need for additional preprocessing.

Under this probabilistic modeling framework, style, content, and noise can be jointly represented within a unified latent space. As illustrated in Figure 3, we begin by constructing the image \mathbf{I} through the spatial concatenation of \mathbf{X} and \mathbf{X}_c . A random binary mask image \mathbf{I}_m is then applied to generate the masked input \mathbf{I}_i . After encoding with the VAE encoder and applying patchification, we obtain the masked image tokens \mathbf{F}_i together with the mask tokens \mathbf{F}_m . Finally, the noisy image tokens \mathbf{F}_n are concatenated with \mathbf{F}_i and \mathbf{F}_m along the channel dimension. The overall procedure

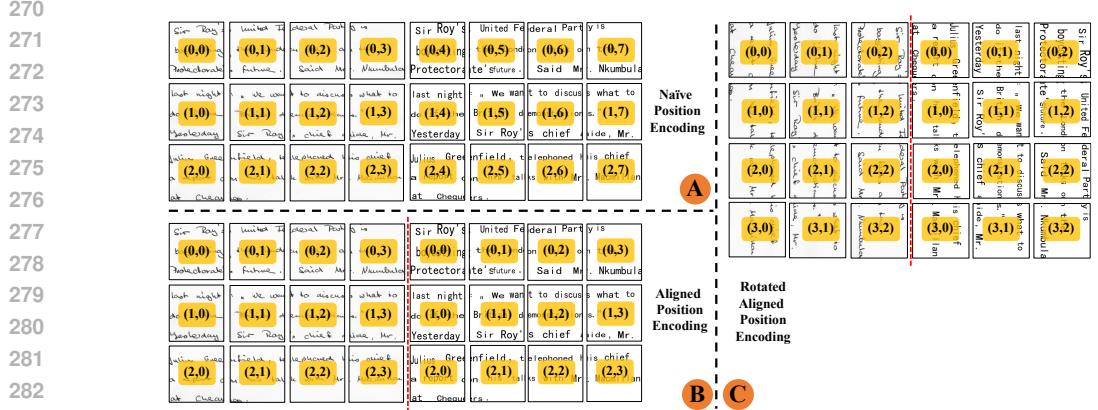


Figure 4: Comparison of different positional encoding strategies. We introduce Aligned Position Encoding (APE) to better guide the spatial layout of generated handwritten text, and further propose a variant, R-APE, tailored for long text lines.

can be formally expressed as follows:

$$\mathbf{I} = \mathbf{X} \odot \mathbf{X}_c, \quad (5)$$

$$\mathbf{I}_i = \mathbf{I} \otimes (\mathbf{1} - \mathbf{I}_m), \quad (6)$$

$$\mathbf{F}_i = \text{Patchify}(\text{VAE}(\mathbf{I}_i)), \quad (7)$$

$$\mathbf{F}_m = \text{Patchify}(\mathbf{I}_m), \quad (8)$$

$$\mathbf{F}_{input} = \mathbf{F}_n \odot \mathbf{F}_i \odot \mathbf{F}_m. \quad (9)$$

Here, \odot denotes the concatenation operation along the spatial dimension, \otimes represents the element-wise (Hadamard) product, \odot denotes the concatenation operation along the channel dimension, \mathbf{F}_i encapsulates the information from both \mathbf{X}_{ctx} and \mathbf{X}_c . Since the same VAE encoder is utilized, style, content, and noise are jointly represented within a unified feature space.

3.3.2 MASKED CONDITIONAL FLOW-MATCHING OBJECTIVE

We employ a flow-matching training objective to optimize the model. Specifically, given a clean latent code \mathbf{x}_0 , a Gaussian noise sample $\mathbf{z}_1 \sim \mathcal{N}(0, I)$, and a time-dependent noise scale σ_t , we generate the noisy latent input \mathbf{x}_0 via a convex combination:

$$\mathbf{x}_t = (1 - \sigma_t)\mathbf{x}_0 + \sigma_t\mathbf{z}_1. \quad (10)$$

The model learns to estimate the velocity vector pointing from \mathbf{x}_0 to \mathbf{z}_1 and the training loss is formulated as:

$$\mathcal{L}_{\text{img-CFM-m}}(\theta) = \mathbb{E}_{t, \mathbf{x}_0, \mathbf{z}_1} \|\mathbf{m} \odot (\hat{\mathbf{v}}_\theta(\mathbf{x}_t, t, \mathbf{c}) - (\mathbf{z}_1 - \mathbf{x}_0))\|_2^2, \quad (11)$$

where $\hat{\mathbf{v}}_\theta$ denotes the model's velocity prediction, \mathbf{m} indicates the masked regions of the latent tokens, and \mathbf{c} comprises \mathbf{X}_{ctx} and \mathbf{X}_c . To ensure a concise optimization process, the training excludes additional objectives such as perceptual loss or CTC loss.

3.3.3 ROTATED ALIGNED POSITION ENCODING

Directly fine-tuning pretrained diffusion transformer models fails to fully exploit their generative capacity for HTG tasks. We therefore redesign positional encodings to better support multi-line handwritten text with arbitrary line lengths. As shown in Figure 4, naïve 2D RoPE arranges tokens row by row in the concatenated image \mathbf{I} , where standard-font and handwritten tokens are interleaved. However, since the length of text-line images varies significantly, the model struggles to distinguish whether a given token should serve as a style condition or a content condition.

To mitigate this issue, we propose Aligned Positional Encoding (APE). In APE, the token arrangement of \mathbf{I} remains unchanged, while the positional encodings assigned to \mathbf{X}_c are directly shared with

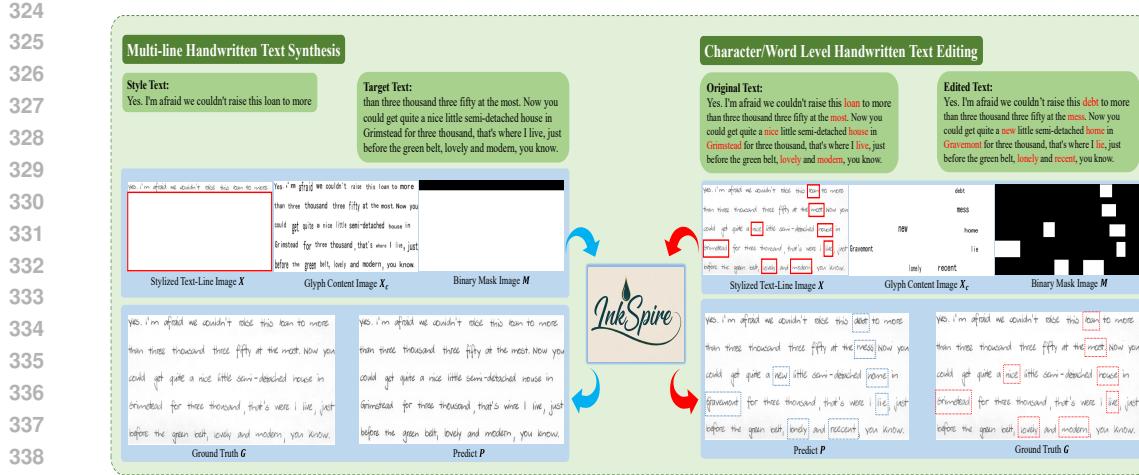


Figure 5: Inference applications of InkSpire, encompassing Multi-line Handwritten Text Synthesis and Character/Word-level Handwritten Text Editing. All applications are guided by the mask image M , the stylized image X_s and the content image X_c .

their counterparts in \mathbf{X} . For cases where text-line images are wider than tall, we further introduce Rotated APE (R-APE): \mathbf{X} and \mathbf{X}_c are rotated 90° clockwise before concatenation, so that target tokens and their content-condition counterparts remain spatially close in the positional space.

3.4 APPLICATIONS OF INFERENCE

Owing to the powerful contextual generation ability of diffusion transformer models and the precise spatial control provided by \mathbf{X}_c , our framework enables versatile inference through simple adjustments to the mask M or modifications to \mathbf{X}_c . The overall procedure is illustrated in Figure 5.

- Multi-line Handwritten Text Synthesis: Given a single style reference image from a writer, previous approaches are typically constrained to generating only one or a few text lines. In contrast, our model is capable of synthesizing an arbitrary number of text lines simultaneously, by placing the style reference in the first line and masking the remaining parts, conditioned on the multi-line content image \mathbf{X}_c .
- Character/Word Level Handwritten Text Editing: By providing a mask that specifies the regions to be edited, together with an edit-content image rendered in standard font, InkSpire can accurately modify multiple words within a handwritten text image while preserving the unmasked regions.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

4.1.1 DATASETS

To validate the effectiveness of InkSpire in synthesizing handwritten text-line images, experiments are conducted on the IAM dataset (Marti & Bunke, 2002) for English and on CASIA-HWDB2.0-2.2 (Liu et al., 2011) and ICDAR2013 datasets (Yin et al., 2013) for Chinese. IAM comprises 13,353 English text-line images, with 496 writers' samples for training and the remaining 161 for testing. For Chinese, CASIA-HWDB2.0-2.2 contains 52,230 text-line images from 1,019 writers for training, while ICDAR2013 includes 3432 text-line images from 60 writers for testing. All datasets provide layout annotations at the word level for English and the character level for Chinese.

378	Text: After what seemed an undue period of repetition	379	Text: 对目前的宏观经济衰退风险起到了主要的推动作用
380	Style Ref <i>went out to observe the landing of the Tibetan</i>	381	Style Ref <i>客观而言，此轮金融风暴源于欧美发达国家</i>
382	VaTr <i>After what seemed an undue period of repetitions</i>	383	One-DM <i>对目前的宏观经济衰退风险起到了主要的推动作用</i>
384	One-DM <i>After what seemed an undue period of repetition</i>	385	TGC-Diff <i>对目前的宏观经济衰退风险起到了主要的推动作用</i>
386	DiffPen <i>After what seemed an undue period of repetition</i>	387	InkSpire <i>对目前的宏观经济衰退风险起到了主要的推动作用</i>
388	InkSpire <i>After what seemed an undue period of repetition</i>	389	GT <i>对目前的宏观经济衰退风险起到了主要的推动作用</i>
390	GT <i>After what seemed an undue period of repetition</i>	391	Text: 姚明悄悄地改变了自己的打法，像是一台“推土机”。
392	Text: I followed, seeing how the sun gilded	393	Style Ref <i>还剩 5 分钟时，姚明摇摇晃晃身后，并没有和以前一样先</i>
394	Style Ref <i>At the school, which is charming, ci-</i>	395	One-DM <i>姚明悄悄地改变了自己的打法，像是一台“推土机”。</i>
396	VaTr <i>I followed, seeing how the sun gilded</i>	397	TGC-Diff <i>姚明悄悄地改变了自己的打法，像是一台“推土机”。</i>
398	One-DM <i>I followed, seeing how the sun gilded</i>	399	InkSpire <i>姚明悄悄地改变了自己的打法，像是一台“推土机”。</i>
400	DiffPen <i>I followed, seeing how the sun gilded</i>	401	GT <i>姚明悄悄地改变了自己的打法，像是一台“推土机”。</i>
402	FLUSH <i>I followed, seeing how the sun gilded</i>	403	Text: 家的豪华并非只是物品的奢侈，还可以是设计尺度上的
404	GT <i>I followed, seeing how the sun gilded</i>	405	Style Ref <i>从威严、厚重走向亲切、自然。这里，体现主人对于家的</i>
406	Text: beginning to find the stereotyped, flattish, happy	407	One-DM <i>豪华的奢侈并非只是物品的奢侈，还可以是设计尺度上的</i>
408	Style Ref <i>After a long wait a bus appeared</i>	409	TGC-Diff <i>豪华的奢侈并非只是物品的奢侈，还可以是设计尺度上的</i>
410	VaTr <i>beginning to find the stereotyped, flattish, happy</i>	411	InkSpire <i>豪华的奢侈并非只是物品的奢侈，还可以是设计尺度上的</i>
412	One-DM <i>beginning to find the stereotyped, flattish, happy</i>	413	GT <i>豪华的奢侈并非只是物品的奢侈，还可以是设计尺度上的</i>
414	DiffPen <i>beginning to find the stereotyped, flattish, happy</i>	415	
416	FLUSH <i>beginning to find the stereotyped, flattish, happy</i>	417	
418	GT <i>beginning to find the stereotyped, flattish, happy</i>	419	
420	4.1.2 EVALUATION METRICS	421	
422	The evaluation metrics encompass two key aspects: style and content. To assess style diversity and	423	
424	consistency, we employ the Fréchet Inception Distance (FID) (Heusel et al., 2017), Kernel Inception	425	
426	Distance (KID) (Bińkowski et al., 2018), and the task-specific Handwriting Distance (HWD) (Pippi	427	
428	et al., 2023b). For content accuracy, we adopt the Correct Rate (CR) and Accuracy Rate (AR) (Yin	428	
429	et al., 2013) for Chinese text, while for English, we use the Absolute Character Error Rate Difference	429	
430	(Δ CER) (Nikolaidou et al., 2024).	430	
431	4.1.3 COMPARED METHODS	431	
432	For English handwritten text line generation, we compare our method against representative models,	432	
433	including HWT (Bhunia et al., 2021), VATr (Vannerle et al., 2024), One-DM (Dai et al., 2024), and	433	
434	DiffPen (Nikolaidou et al., 2024). For Chinese handwritten text line generation, the comparison is	434	
435	conducted with One-DM and TGC-Diff (Wang et al., 2025).	435	
436	4.1.4 IMPLEMENTATION DETAILS	436	
437	For layout generation, we employ 10 transformer layers across all three modeling approaches. During	437	
438	inference, the three layout modeling strategies differ in how layouts are produced: in the autoregressive	438	
439	model, the reference layout is encoded as prefix tokens and the target layout is generated	439	
440	sequentially in a token-by-token manner; in the masked layout model, the input contains an observed	440	
441	reference layout and a masked target layout, and all masked tokens are predicted in a single forward	441	

Method	IAM Layout				ICDAR2013 Layout			
	$\Delta x \downarrow$	$\Delta y \downarrow$	$\Delta w \downarrow$	$\Delta h \downarrow$	$\Delta x \downarrow$	$\Delta y \downarrow$	$\Delta w \downarrow$	$\Delta h \downarrow$
Autoregressive Modeling	5.60	17.04	11.42	13.67	7.28	19.25	14.29	12.88
Masked Layout Modeling	5.18	14.51	6.52	8.04	6.42	16.71	13.67	10.56
Masked Modeling with CFM	4.74	14.39	4.74	4.94	5.13	14.85	12.43	8.74

Table 1: Layout prediction results on IAM and ICDAR2013 datasets. Values in the table correspond to \mathcal{L}_1 losses multiplied by 10^3 ; lower values indicate better performance (\downarrow).

	FID \downarrow	KID \downarrow	HWD \downarrow	ΔCER \downarrow		FID \downarrow	KID \downarrow	HWD \downarrow	CR \uparrow	AR \uparrow
HWT	44.72	43.49	2.97	0.33						
VATr	34.00	29.68	2.38	0.03						
One-DM	43.89	44.48	2.83	0.13						
DiffPen	12.89	9.73	2.13	0.03						
InkSpire	7.92	4.83	0.62	0.01						

Table 2: English text line generation results on IAM. The KID is multiplied by 10^3 . Lower is better (\downarrow).

Table 3: Chinese text line generation results on ICDAR2013. The KID is multiplied by 10^3 . Lower is better (\downarrow), higher is better (\uparrow).

pass; in the Conditional Flow Matching (CFM) model, the same masked positions as in the masked layout model are used, but the masked tokens are generated through a continuous denoising process solved with a 10-step ODE solver, enabling smooth and flexible layout prediction.

For image generation, our approach builds upon the pre-trained FLUX.1-Fill-dev model (Labs (2024)), a latent rectified flow transformer tailored for image editing. To adapt to multilingual generation, training samples across languages are integrated within a unified training pipeline. The image patch size is set to $P = 1024$, and the batch size is 4. The model is optimized using the Prodigy optimizer with an initial learning rate of 1 and a weight decay of 0.01. Fine-tuning is performed via Low-Rank Adaptation (LoRA) with a rank of 32 and a LoRA scaling factor of 32, trained for 20,000 iterations on four A100 GPUs (40 GB). In total, the LoRA modules introduce approximately 115.9M trainable parameters. A comprehensive description of all fine-tuned LoRA parameters can be found in Appendix Table 8. The fine-tuned model achieves high-quality synthesis with about 20 ODE denoising steps. For more implementation details of layout generation and image generation, please refer to Appendix A.3.

4.2 ADDITIONAL ANALYSIS OF STYLIZED LAYOUT GENERATION

We compare three layout modeling strategies—Autoregressive, Masked, and Masked with Conditional Flow Matching (CFM)—using average losses of four layout-specific features. For all methods, the tokens corresponding to the first line of each paragraph are provided as reference tokens, and the models are tasked with predicting the layout parameters of all remaining tokens. Under this setting, the masked layout model and CFM apply masking to all non-reference tokens. The evaluation metric is computed by averaging the L1 deviations over the full set of tokens that require prediction. As shown in Table 1, CFM consistently outperforms the others, demonstrating its effectiveness in capturing complex spatial dependencies while preserving layout coherence.

4.3 STYLIZED HANDWRITTEN TEXT-LINE GENERATION

Although InkSpire is trained on multi-line images to capture in-context stylistic features, we follow the one-shot evaluation protocol, where only a single reference text-line is provided. The goal is to generate new lines with the same style but different content. During inference, we consistently select references by using the second line to generate the first, and then the first line for the remaining ones, which mitigates intra-writer style variations across paragraphs.

	FID \downarrow	KID \downarrow	HWD \downarrow	ΔCER \downarrow
baseline	15.12	19.27	0.97	0.11
+APE	9.31	7.21	0.58	0.05
+R-APE	7.92	4.83	0.62	0.01

Table 4: Ablation study of positional encoding on the IAM Lines dataset. The KID is multiplied by 10^3 . Lower is better (\downarrow).

	FID \downarrow	KID \downarrow	HWD \downarrow	ΔCER \downarrow
F-TopMask	8.73	6.13	0.78	0.07
R-Mask	7.92	4.83	0.62	0.01

Table 6: Ablation study of masking strategies on the IAM Lines dataset. The KID is multiplied by 10^3 . Lower is better (\downarrow).

	FID \downarrow	KID \downarrow	HWD \downarrow	CR \uparrow	AR \uparrow
baseline	17.57	18.61	0.53	89.72	88.91
+APE	11.75	12.35	0.42	91.87	90.63
+R-APE	10.98	11.45	0.41	92.92	91.56

Table 5: Ablation study of positional encoding on the ICDAR2013 Lines dataset. The KID is multiplied by 10^3 . Lower is better (\downarrow), higher is better (\uparrow).

	FID \downarrow	KID \downarrow	HWD \downarrow	CR \uparrow	AR \uparrow
F-TopMask	11.57	13.41	0.48	92.48	91.34
R-Mask	10.98	11.45	0.41	92.92	91.56

Table 7: Ablation study of masking strategies on the ICDAR2013 Lines dataset. The KID is multiplied by 10^3 . Lower is better (\downarrow), higher is better (\uparrow).

Quantitative results, as presented in Table 2 and Table 3, demonstrate that InkSpire consistently outperforms state-of-the-art methods on both English and Chinese datasets. As illustrated in Figure 6, our model not only produces handwritten styles that are more visually aligned with the reference, but also exhibits significantly improved character structure accuracy. These findings provide strong evidence of the superiority of our approach and are well-aligned with the quantitative metrics.

4.4 ABLATION STUDY

4.4.1 ABLATION STUDY ON POSITION ENCODING

From Table 4 and Table 5, we observe that Aligned Positional Encoding (APE) markedly enhances style diversity and content accuracy by helping the model better distinguish style from content tokens. While naïve positional encoding can handle single-line generation, it often copies input images in multi-line settings and is sensitive to resolution. Moreover, the rotated variant (R-APE) yields further gains on both English and Chinese datasets, as it better localizes tokens in long text lines under the one-shot setting.

4.4.2 ABLATION STUDY ON MULTI-LINE MASKED INFILLING STRATEGY

In addition to the Random-Size Multi-Region Masking (R-Mask) strategy described in Section 3.3.1, we introduce a Fixed Top-Region Unmasked Masking (F-TopMask) scheme as an ablation variant. Since masking is performed on a 1024×1024 image patch, the fixed-mask design keeps only the top 128×1024 region visible while masking all remaining lower areas. This setup resembles the inference scenario where only the first text line is provided as a style reference and the model must generate all subsequent lines. The quantitative results in Table 6 and Table 7 demonstrate that our R-Mask strategy yields moderate improvements over the fixed-region alternative.

5 CONCLUSION

We introduced InkSpire, a diffusion transformer that unifies style, content, and noise for handwritten text generation. By removing explicit encoders and adopting multi-line masked infilling with revised positional encoding, it enables efficient training, arbitrary-length synthesis, and fine-grained editing. Trained on English and Chinese corpora, InkSpire outperforms prior methods in fidelity and stylistic diversity. Future work will extend to more languages and datasets to enhance generalization.

540 REFERENCES
541

542 Eloi Alonso, Bastien Moysset, and Ronaldo Messina. Adversarial generation of handwritten text
543 images conditioned on sequences. In *International Conference on Document Analysis and Recog-*
544 *nition*, pp. 481–486, 2019.

545 Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Fahad Shahbaz
546 Khan, and Mubarak Shah. Handwriting transformers. In *Proceedings of the IEEE/CVF interna-*
547 *tional conference on computer vision*, pp. 1086–1094, 2021.

548 Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd
549 gans. *arXiv preprint arXiv:1801.01401*, 2018.

550 Frederic Boesel and Robin Rombach. Improving image editing models with generative data refine-
551 ment. In *The Second Tiny Papers Track at ICLR 2024*, 2024.

552 Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image
553 editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern*
554 *recognition*, pp. 18392–18402, 2023.

555 Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng
556 Zhang, Fengbin Gao, Peihan Xu, et al. Hidream-i1: A high-efficient image generative foundation
557 model with sparse diffusion transformer. *arXiv preprint arXiv:2505.22705*, 2025.

558 Haoxing Chen, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Changhua Meng, Huijia Zhu, Weiqiang
559 Wang, et al. Diffute: Universal text editing diffusion model. *Advances in Neural Information*
560 *Processing Systems*, 36:63062–63074, 2023.

561 Gang Dai, Yifan Zhang, Qingfeng Wang, Qing Du, Zhuliang Yu, Zhuoman Liu, and Shuangping
562 Huang. Disentangling writer and character styles for handwriting generation. In *Proceedings of*
563 *the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5977–5986, 2023.

564 Gang Dai, Yifan Zhang, Quhui Ke, Qiangya Guo, and Shuangping Huang. One-dm: One-shot
565 diffusion mimicker for handwritten text generation. *European Conference on Computer Vision*,
566 2024.

567 Haisong Ding, Bozhi Luan, Dongnan Gui, Kai Chen, and Qiang Huo. Improving handwritten ocr
568 with training samples generated by glyph conditional denoising diffusion probabilistic model. In
569 *International Conference on Document Analysis and Recognition*, pp. 20–37, 2023.

570 Ji Gan and Weiqiang Wang. HiGAN: Handwriting imitation conditioned on arbitrary-length texts
571 and disentangled styles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol-
572 ume 35, pp. 7484–7492, 2021.

573 Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter.
574 Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in*
575 *neural information processing systems*, 30, 2017.

576 Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong
577 Feng, Yu Liu, and Jingren Zhou. In-context lora for diffusion transformers. *arXiv preprint*
578 *arXiv:2410.23775*, 2024.

579 Yuxin Kong, Canjie Luo, Weihong Ma, Qiyuan Zhu, Shenggao Zhu, Nicholas Yuan, and Lianwen
580 Jin. Look closer to supervise better: one-shot font generation via component-based discriminator.
581 In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.
582 13482–13491, 2022.

583 Atsunobu Kotani, Stefanie Tellex, and James Tompkin. Generating handwriting via decoupled style
584 descriptors. In *European Conference on Computer Vision*, pp. 764–780. Springer, 2020.

585 Black Forest Labs. Flux: Official inference repository for flux.1 models, 2024. URL <https://github.com/black-forest-labs/flux>. Accessed: 2024-11-12.

594 Cheng-Lin Liu, Fei Yin, Da-Han Wang, and Qiu-Feng Wang. Casia online and offline chinese
 595 handwriting databases. In *International Conference on Document Analysis and Recognition*, pp.
 596 37–41, 2011.

597 Wei Liu, Fangyue Liu, Fei Ding, Qian He, and Zili Yi. Xmp-font: self-supervised cross-modality
 598 pre-training for few-shot font generation. In *Proceedings of the IEEE/CVF Conference on Com-*
 599 *puter Vision and Pattern Recognition*, pp. 7905–7914, 2022.

600 Troy Luhman and Eric Luhman. Diffusion models for handwriting generation. *arXiv preprint*
 601 *arXiv:2011.06704*, 2020.

602 U-V Marti and Horst Bunke. The IAM-database: an English sentence database for offline handwrit-
 603 ing recognition. pp. 39–46, 2002.

604 Konstantina Nikolaïdou, George Retsinas, Vincent Christlein, Mathias Seuret, Giorgos Sfikas,
 605 Elisa Barney Smith, Hamam Mokayed, and Marcus Liwicki. Wordstylist: Styled verbatim hand-
 606 written text generation with latent diffusion models. In *International Conference on Document*
 607 *Analysis and Recognition*, pp. 384–401, 2023.

608 Konstantina Nikolaïdou, George Retsinas, Giorgos Sfikas, and Marcus Liwicki. Diffusionpen: To-
 609 wards controlling the style of handwritten text generation. In *European Conference on Computer*
 610 *Vision*, pp. 417–434. Springer, 2024.

611 Vittorio Pippi, Silvia Cascianelli, and Rita Cucchiara. Handwritten text generation from visual
 612 archetypes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog-*
 613 *nition*, pp. 22458–22467, 2023a.

614 Vittorio Pippi, Fabio Quattrini, Silvia Cascianelli, and Rita Cucchiara. Hwd: A novel evaluation
 615 score for styled handwritten text generation. *arXiv preprint arXiv:2310.20316*, 2023b.

616 Vittorio Pippi, Fabio Quattrini, Silvia Cascianelli, Alessio Tonioni, and Rita Cucchiara. Zero-shot
 617 styled text image generation, but make it autoregressive. In *Proceedings of the Computer Vision*
 618 *and Pattern Recognition Conference*, pp. 7910–7919, 2025.

619 Min-Si Ren, Yan-Ming Zhang, Qiu-Feng Wang, Fei Yin, and Cheng-Lin Liu. Diff-writer: A dif-
 620 fusion model-based stylized online handwritten chinese character generator. In *International*
 621 *Conference on Neural Information Processing*, pp. 86–100. Springer, 2023.

622 Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh,
 623 and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *Pro-*
 624 *ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8871–
 625 8879, 2024.

626 Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha
 627 Zhang, and Mohit Bansal. Unifying vision, text, and layout for universal document processing.
 628 In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp.
 629 19254–19264, 2023.

630 Bram Vanherle, Vittorio Pippi, Silvia Cascianelli, Nick Michiels, Frank Van Reeth, and Rita Cuc-
 631 chiara. Vatr++: Choose your words wisely for handwritten text generation. *IEEE Transactions*
 632 *on Pattern Analysis and Machine Intelligence*, 2024.

633 Honglie Wang, Minsi Ren, Yan-Ming Zhang, Fei Yin, and Cheng-Lin Liu. Template-guided cas-
 634 caded diffusion for stylized handwritten chinese text-line generation. In *Document Analysis and*
 635 *Recognition – ICDAR 2025*, pp. 149–166, Cham, 2025. Springer Nature Switzerland.

636 Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li,
 637 Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceed-*
 638 *ings of the Computer Vision and Pattern Recognition Conference*, pp. 13294–13304, 2025.

639 Yangchen Xie, Xinyuan Chen, Li Sun, and Yue Lu. DG-Font: Deformable generative networks for
 640 unsupervised font generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision*
 641 *and Pattern Recognition*, pp. 5130–5140, 2021.

648 Gang Yao, Liangrui Peng, Zhiyu Li, Tianqi Zhao, Kemeng Zhao, Ning Ding, and Yao Tao. Multi-
649 level style control for chinese handwriting generation: G. yao et al. *International Journal on*
650 *Document Analysis and Recognition (IJDAR)*, pp. 1–12, 2025.
651

652 Fei Yin, Qiu-Feng Wang, Xu-Yao Zhang, and Cheng-Lin Liu. ICDAR 2013 chinese handwriting
653 recognition competition. In *International Conference on Document Analysis and Recognition*,
654 pp. 1464–1470, 2013.

655 Xu-Yao Zhang, Fei Yin, Yan-Ming Zhang, Cheng-Lin Liu, and Yoshua Bengio. Drawing and recog-
656 nizing chinese characters with recurrent neural network. *IEEE Transactions on Pattern Analysis*
657 and *Machine Intelligence*, 40(4):849–862, 2017.

658 Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. In-context edit: Enabling instructional
659 image editing with in-context generation in large scale diffusion transformer. *arXiv preprint*
660 *arXiv:2504.20690*, 2025.
661

662 Bocheng Zhao, Jianhua Tao, Minghao Yang, Zhengkun Tian, Cunhang Fan, and Ye Bai. Deep
663 imitator: Handwriting calligraphy imitation via deep attention networks. *Pattern Recognition*,
664 104:107080, 2020.

665 Yuanzhi Zhu, Zhaohai Li, Tianwei Wang, Mengchao He, and Cong Yao. Conditional text image
666 generation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer*
667 *Vision and Pattern Recognition*, pp. 14235–14245, 2023.
668

669

670

671

672

673

674

675

676

677

678

679

680

681

682

683

684

685

686

687

688

689

690

691

692

693

694

695

696

697

698

699

700

701

702

A APPENDIX

703

A.1 INFORMATION ABOUT USE OF AI ASSISTANTS

704 The authors used AI-assisted tools solely for language polishing, including grammar, spelling, and
 705 expression refinement. All conceptual contributions, technical innovations, experimental design,
 706 and analyses presented in this work were independently performed by the authors without reliance
 707 on AI for content generation or scientific reasoning.

710

A.2 USER STUDIES

711

A.2.1 USER PREFERENCE STUDY

712 We conduct a human evaluation to assess the perceptual quality of synthesized Chinese handwritten
 713 text-line images, focusing on style fidelity and content correctness. Participants, all of whom hold
 714 postgraduate-level education, compare the outputs of our method with two state-of-the-art baselines,
 715 One-DM and TGC-Diff. In each trial, a writer is randomly sampled from the ICDAR2013 dataset,
 716 and one of their handwritten text-line images is provided as a style reference, together with an
 717 identical content prompt for all methods. Participants are shown the reference text-line alongside
 718 multiple candidate images generated by the three models and are asked to select the sample with the
 719 highest overall generation quality. Figure 7 illustrates an example of the questionnaire instructions
 720 and its corresponding question items. The evaluation consists of 30 rounds, yielding 900 valid
 721 responses from 30 volunteers. As shown in the Figure 8(a), our method receives the highest number
 722 of user selections, indicating its superior perceptual quality in handwritten text generation.

723 **Instructions**
 724 This survey contains **30 evaluation questions**.
 725 For each question, you will be given a **target text** and a **style reference image**.
 726 Options **A, B, and C** show images generated by **three different models** (in randomized order).
 727 Your task is to **select the image with the highest generation quality**.

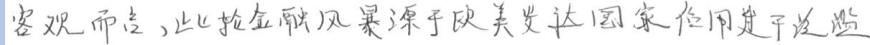
728 “**Generation quality**” primarily includes two aspects:
 729 **1. Style:** How similar the generated handwriting style is to the reference
 730 (e.g., stroke thickness, curvature, connections, and other local stylistic details).
 731 **2. Content:** The correctness and clarity of the generated characters.
 732 **Note:** If the reference writing style is exaggerated, easily distorted, or blurry, please prioritize style
 733 consistency over content accuracy. Please ignore occasional hallucinated content that exceeds the given
 734 target text.

735 In short, choose the image that **most closely resembles being written by the same person** as in the
 736 style reference image. For best results, please complete the survey **on a computer** and **click to zoom in**
 737 on each image before making a selection.

738 **要生成的文本内容**

739 退风险起到了主要的推动作用,且其本身仅仅是全球宏观

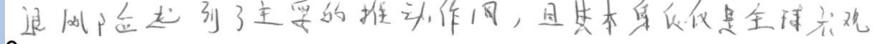
740 **风格参考图片**

741 

742 **A.**

743 

744 **B.**

745 

746 **C.**

747 

748 **Figure 7: User Preference Study Instructions.**

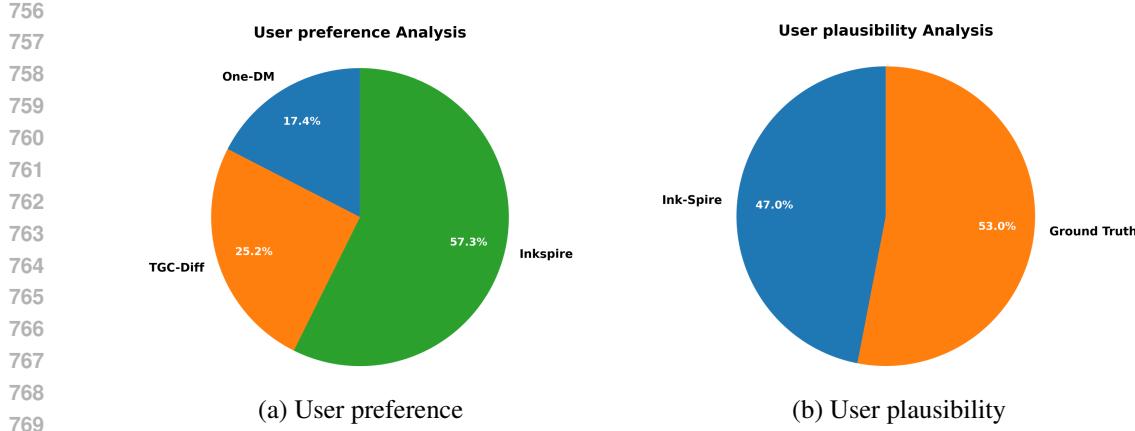


Figure 8: Overall caption describing both images.

A.2.2 USER PLAUSIBILITY STUDY

We conduct a user plausibility study to evaluate whether text-line images generated by InkSpire are perceptually indistinguishable from real handwriting. Participants are first presented with 30 authentic handwritten text-line samples, which serve as style reference images. In each question, they are then shown two candidate images: one genuine sample written by the same author and one generated by our model. Their task is to determine which image appears more likely to be written by the same writer as the reference. Figure 9 illustrates an example of the questionnaire instructions and its corresponding question items. A total of 23 participants provide 690 valid responses. As shown in Figure 8(b), the selection accuracy converges to approximately, indicating performance at chance level. This suggests that the text-line images produced by InkSpire are nearly indistinguishable from real handwriting.

Instructions

This survey contains **30 evaluation questions**.

For each question, you will be given a **target text** and a **style reference image**.

In this study, **only two candidate images** are presented.

Please select the image that **most closely resembles the style of the reference image** (i.e., appears to be written by the same author). The order of the two options has been randomized.

Please **ignore occasional hallucinated content** that exceeds the specified text.

For best results, complete the survey **on a computer** and **click to zoom in** on each image before making your selection.

要生成的文本内容

9月份想成为莱德杯欧洲队的一员而去美国肯塔基州的路易斯维尔

风格参考图片

丹尼尔-希普拉在1月6日夺得梅塞德斯锦标赛冠军之后或许可以实现

A.

9月份想成为莱德杯欧洲队的一员而去美国肯塔基州的路易斯维尔

B.

9月份想成为莱德杯欧洲队的一员而去美国肯塔基州的路易斯维尔，

Figure 9: User Plausibility Study Instructions.

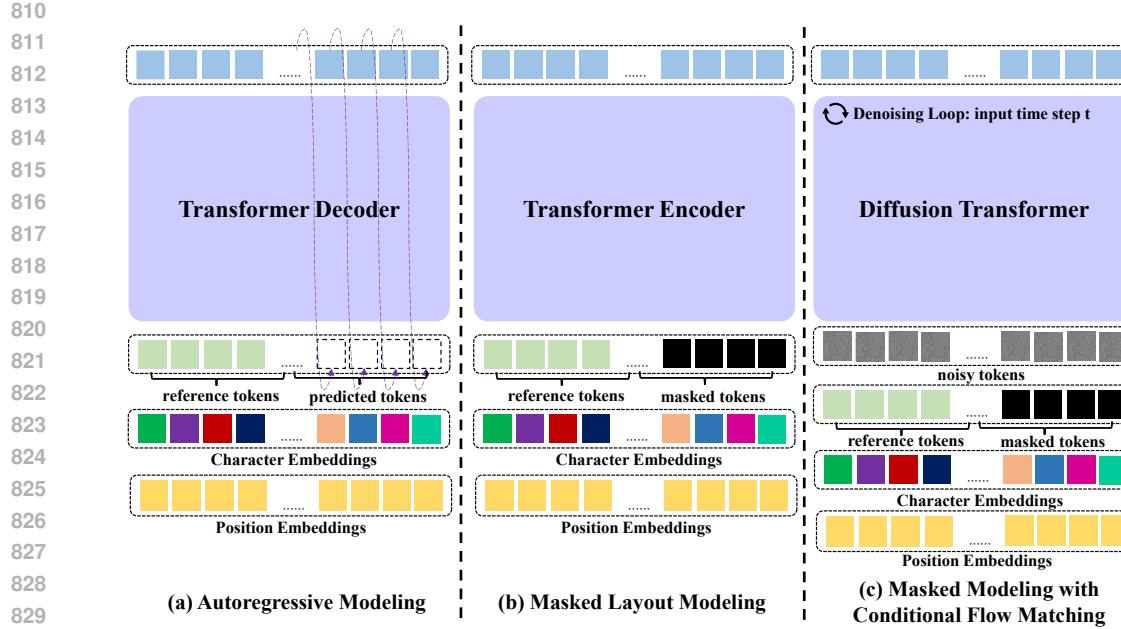


Figure 10: Layout Inference Architectures.

A.3 MORE IMPLEMENTATION DETAILS

A.3.1 LAYOUT GENERATION DETAILS

We provide a detailed description of the three layout generation strategies discussed in the main text:

- Autoregressive Modeling: This approach predicts the layout of each target token using an autoregressive Transformer architecture. The model input consists of the embedding of the current character (or, for English words, the sum of the individual character embeddings to form a word-level embedding), positional embeddings, and the layout information of preceding tokens. We employ 10 Transformer decoder layers and optimize the model using an L1 regression loss.
- Masked Layout Modeling: In this strategy, the input consists of character embeddings, positional embeddings, and a partially masked layout representation of tokens. We adopt two masking strategies: (1) randomly masking a contiguous segment of arbitrary length, and (2) masking each token independently with a 20% probability. Each masking strategy is applied with equal probability (50%). The model is trained to predict the layout of the masked tokens. We employ 10 Transformer encoder layers and optimize the model using an L1 regression loss computed over the masked tokens.
- Masked Modeling with Conditional Flow Matching: This variant adopts the same masking strategies as Masked Layout Modeling, but extends the masked modeling framework by incorporating a time-dependent denoising condition inspired by flow-matching objectives. In addition to character embeddings, positional embeddings, and conditional tokens, all masked tokens are progressively denoised from noisy tokens over multiple timesteps. We employ a 10-layer diffusion Transformer to model this process and optimize it using an L1 loss over the predicted layout values (v-prediction) for the masked tokens.

During inference, we take the tokens of the first sentence in a paragraph as reference tokens and ask the models to predict the remaining tokens. The detailed model structures are illustrated in Figure 10. All Transformer blocks have a hidden dimension of 512 and employ 8 attention heads. The models are trained with a batch size of 160 using the AdamW optimizer with a base learning rate of 1×10^{-4} .

864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895

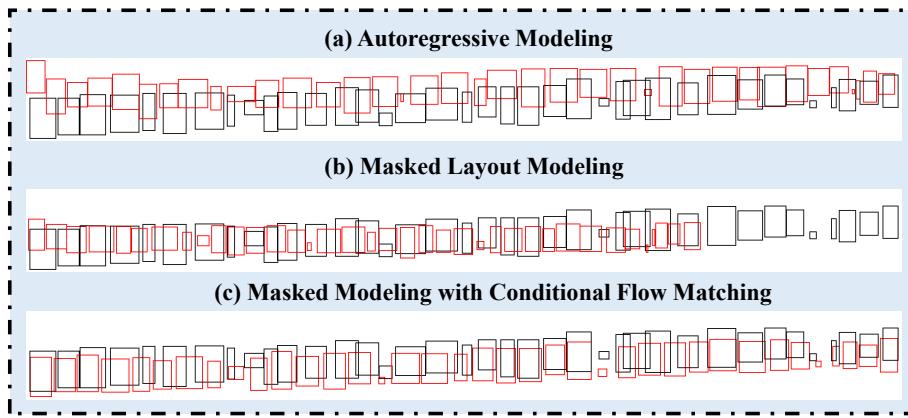


Figure 11: Visualization results of Chinese text-line generation. Black rectangles indicate the target layout, while red rectangles denote the predicted bounding boxes.

896
897
898
899
900

A.3.2 HANDWRITTEN TEXT IMAGE GENERATION DETAILS

We list all LoRA fine-tuning parameters in Table 8. With a LoRA rank of 32 and a scaling factor of 32, the LoRA modules collectively introduce approximately 115.9M trainable parameters.

901
902
903

A.4 ADDITIONAL ANALYSIS OF STYLIZED LAYOUT GENERATION

904
905

We provide two visualization examples for Chinese and English layout generation in Figure 11 and Figure 12, respectively. In each figure, black rectangles denote the target layout, while red rectangles represent the predicted character bounding boxes produced by each method.

906
907
908

In Figure 11, the Chinese example illustrates that our masked modeling with CFM successfully captures the upward-slanting trajectory of the text line—from the lower left toward the upper right. In contrast, the slant is much less pronounced in the layouts generated by the Autoregressive Model and the Masked Layout Modeling baseline.

909
910
911
912

In Figure 12, the English example shows that masked modeling with CFM provides noticeably better control over word spacing and word size compared with the other two strategies, leading to a more coherent and visually consistent layout.

913
914
915
916

Overall, the layouts generated by masked modeling with CFM demonstrate superior visual fidelity, aligning well with the quantitative improvements reported in Table 1.

Parameter Name	#Params (M)
x_embedder	0.1106
transformer_blocks.0-18.norm1.linear	13.072
transformer_blocks.0-18.attn.to_q	3.7355
transformer_blocks.0-18.attn.to_k	3.7355
transformer_blocks.0-18.attn.to_v	3.7355
transformer_blocks.0-18.attn.to_out.0	3.7355
transformer_blocks.0-18.ff.net.2	9.3389
single_transformer_blocks.0-37.norm.linear	14.942
single_transformer_blocks.0-37.proj_mlp	18.678
single_transformer_blocks.0-37.proj_out	22.413
single_transformer_blocks.0-37.attn.to_q	7.4711
single_transformer_blocks.0-37.attn.to_k	7.4711
single_transformer_blocks.0-37.attn.to_v	7.4711
Total	115.91

Table 8: List of all LoRA parameter names.

Model	Style Score (%)
English Models	
HWT	39.62
VATr	45.19
One-DM	50.73
DiffPen	61.26
InkSpire	78.58
Chinese Models	
One-DM	51.34
TGC-Diff	60.46
InkSpire	86.28

Table 9: Style Scores (%) of different models for English and Chinese text-line generation.

936 A.5 ADDITIONAL ANALYSIS OF STYLIZED HANDWRITTEN TEXT-LINE GENERATION

937 A.5.1 VISUALIZATION ANALYSIS OF POSITION ENCODING ABLATION STUDY

938 We provide visualization results for the ablation study on positional encoding. Figure 13 presents
939 examples of both Chinese and English text-line generation. The red bounding boxes highlight in-
940 stances where the model produces structurally incorrect handwriting.

941 From the visualizations, we observe that using the original positional encoding of the Flux-fill back-
942 bone leads to inaccurate spatial localization, causing the model to generate handwriting fragments
943 at incorrect positions within the image. In contrast, APE improves positional consistency, and the
944 proposed R-APE further enhances structural accuracy. The superiority of R-APE over APE is par-
945 ticularly evident in the correctness of character shapes and alignment, which is consistent with the
946 quantitative improvements reported in Table 4 and Table 5.

947 A.5.2 VISUALIZATION ANALYSIS OF MASKED INFILLING STRATEGY ABLATION STUDY

948 We provide visualization results of the ablation study on masked infilling strategy. Figure 14 presents
949 examples of both Chinese and English text-line generation. In both figures, the red bounding boxes
950 highlight cases where the model produces structurally incorrect handwriting.

951 Overall, the stylistic appearance produced by both the random masking (R-Mask) and the fixed
952 top-region masking (F-TopMask) strategies remains comparable and closely aligned with the target
953 style. However, subtle differences emerge at the fine-grained character-structure level: the R-Mask
954 strategy leads to fewer structural errors, likely because its randomly sampled masks expose the
955 model to a wider range of local character patterns during training. These qualitative observations
956 are consistent with the quantitative improvements reported in Table 6 and Table 7.

957 A.5.3 STYLE SCORE EVALUATION

958 We evaluate style fidelity using text-line style classifiers built upon ImageNet-pretrained Swin Trans-
959 formers (about 86.8M parameters), trained separately for English and Chinese. All text-line images
960 are normalized to 64×1024 resolution by proportional resizing to height 64 followed by width-wise
961 cropping or white padding.

962 The resulting classifiers reach 94% accuracy on the 60-way Chinese validation set and 91% on the
963 161-way English set. Their prediction accuracy on generated text lines is reported as the *style score*,
964 indicating how closely a model replicates the target handwriting style.

965 As shown in Table 9, InkSpire consistently attains the highest style scores across both English and
966 Chinese settings, outperforming all baselines and aligning with trends observed in the user study
967 and earlier style-consistency analyses.

972

973

974

975

976

977

978

979

980

981

982

983

984

985

986

987

988

989

990

991

992

993

994

995

996

997

998

999

1000

1001

1002

1003

1004

1005

1006

1007

1008

1009

1010

1011

1012

1013

1014

1015

1016

1017

1018

1019

1020

1021

1022

1023

1024

1025

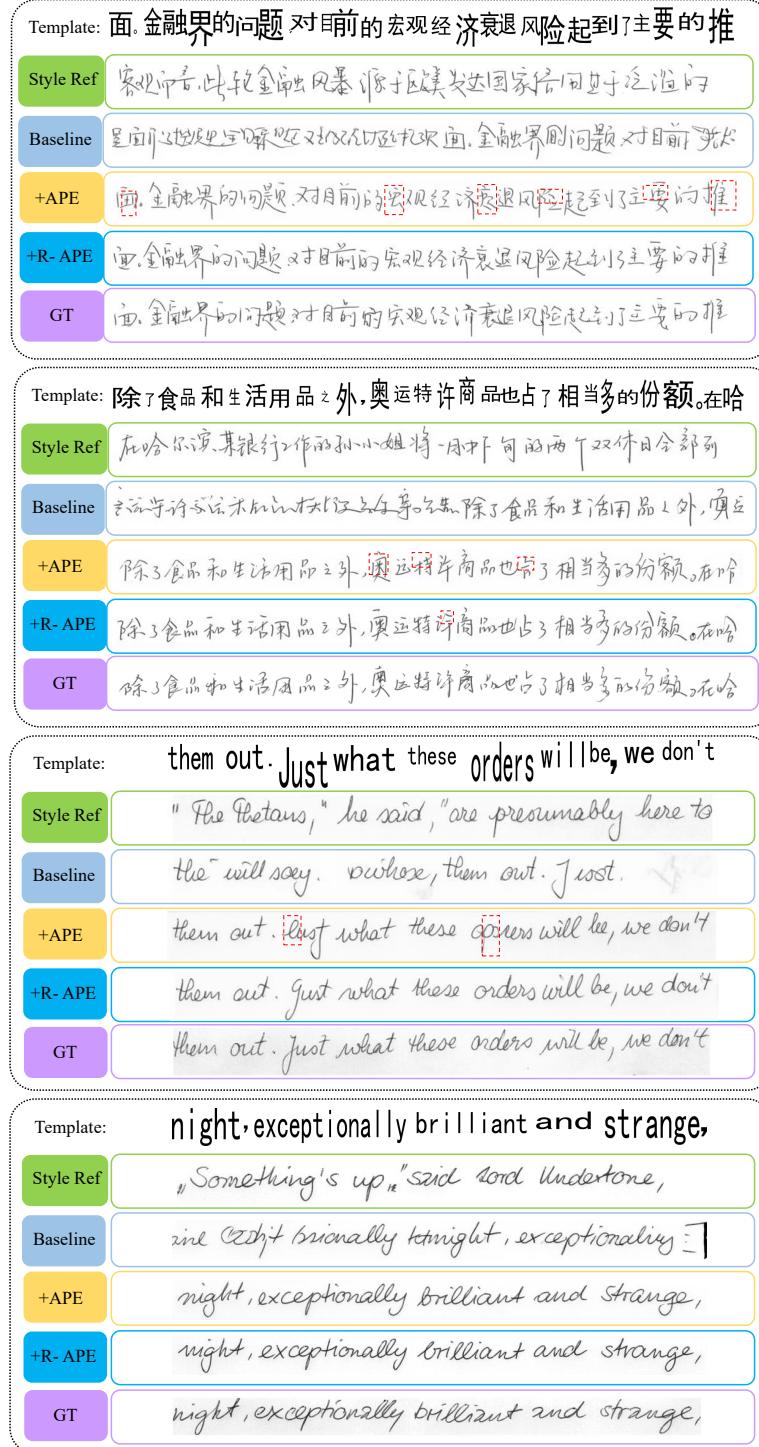


Figure 13: Visualization results of Chinese text-line generation for the positional encoding ablation. Red rectangles highlight cases where incorrect character structures are produced.

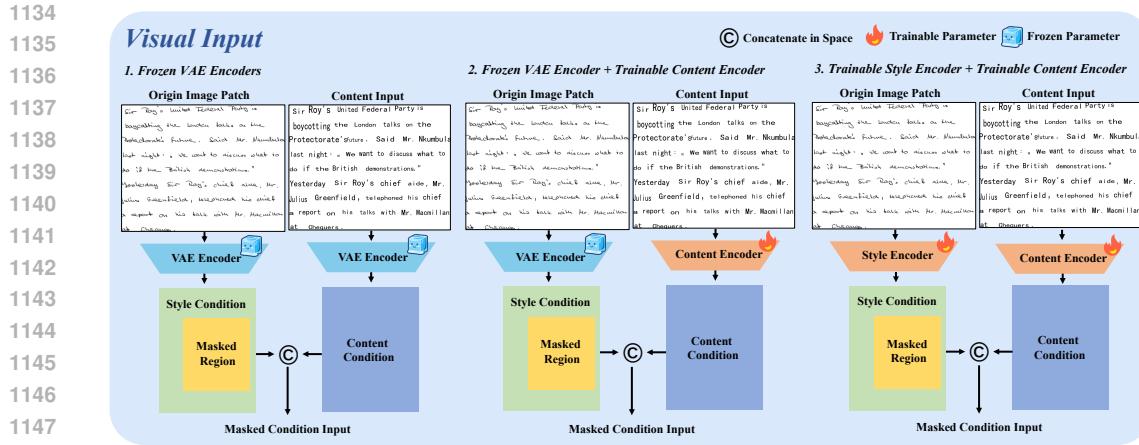


Figure 17: Overview of the visual input processing under different encoder configurations.

	FID ↓	KID ↓	HWD ↓	ΔCER ↓
Fixed-VAE	7.92	4.83	0.62	0.01
C-Enc	20.48	10.31	1.58	0.05
C+S-Enc	17.39	8.97	0.71	0.02

	FID ↓	KID ↓	HWD ↓	CR ↑	AR ↑
Fixed-VAE	10.98	11.45	0.41	92.92	91.56
C-Enc	17.46	19.52	0.91	59.32	55.58
C+S-Enc	16.53	14.37	0.59	94.74	93.48

Table 10: Ablation of encoder configurations on the IAM Lines dataset. C-Enc: only Content Encoder trained from scratch; C+S-Enc: both Content and Style Encoders trained from scratch. KID is scaled by 10^3 .Table 11: Ablation of encoder configurations on the ICDAR2013 Lines dataset. C-Enc: only Content Encoder trained from scratch; C+S-Enc: both Content and Style Encoders trained from scratch. KID is scaled by 10^3 .

A.5.5 IMPACT OF ENCODER DESIGN CHOICES

We conduct an ablation study to analyze the impact of different encoder configurations in our unified encoder-less framework. As illustrated in Figure 17, we design three experimental settings:

- **Pretrained-VAE Encoding (Default Setting).** In the original setup, a pretrained VAE is used to encode both the origin image patch and the content image.
- **Learned Content Encoder Only (C-Enc).** In this variant, we train a content encoder from scratch to extract content features, while the origin image patch continues to be encoded using the frozen pretrained VAE.
- **Jointly Learned Content and Style Encoders (C+S-Enc).** In the third configuration, both the content encoder and the style encoder are jointly learned from scratch.

Following Dai et al. (2024), both the content and style encoders adopt a CNN+Transformer architecture. The CNN backbone is a ResNet-50 pretrained on ImageNet, followed by a Transformer encoder with 3 layers, a hidden dimension of 2048, and 16 attention heads. To adapt the representation to the DiT input format, we append a linear projection layer that maps 2048-dimensional features to 64 dimensions. The proposed encoder contains approximately 124.4M parameters, while the frozen VAE encoder has about 34.3M parameters. Quantitative results for the three configurations are reported in Table 10 and Table 11.

Figure 18 presents visualization results for Chinese and English text-lines. We observe that when training encoders from scratch (both C-Enc and C+S-Enc), the generated strokes tend to appear noticeably lighter. This is likely caused by a distributional bias between the latent space learned from scratch and the latent space of the pretrained VAE. Furthermore, for the C-Enc setting, the model more easily captures complex cursive styles but struggles to maintain accurate character structures. In contrast, the C+S-Enc setting produces outputs that resemble cleaner, more printed-like handwriting, but with reduced stylistic diversity. This suggests that jointly learning content and style encoders may shift the model toward a different balance between content fidelity and stylistic variation. No

1188

1189

1190

1191

1192

1193

1194

1195

1196

1197

1198

1199

1200

1201

1202

1203

1204

1205

1206

1207

1208

1209

1210

1211

1212

1213

1214

1215

1216

1217

1218

1219

1220

1221

1222

1223

1224

1225

1226

1227

1228

1229

1230

1231

1232

1233

1234

1235

1236

1237

1238

1239

1240

1241

Template: 金融衍生品的“创新”走到了华尔街自身无法控制的局面。金融界的	
Style Ref	客观而言，此轮金融风暴源于欧美发达国家信用链于次贷危机的金融界的
Fixed VAE	金融衍生品的“创新”走到了华尔街自身无法控制的局面。金融界的
C-Enc	金融衍生品的“创新”走到了华尔街自身无法控制的局面。金融界的
C+S-Enc	金融衍生品的“创新”走到了华尔街自身无法控制的局面。金融界的
GT	金融衍生品的“创新”走到了华尔街自身无法控制的局面。金融界的
Template: 好航班延误情况下的航班保障工作，白云机场客运部昨日正式启动航	
Style Ref	截止到昨日下午6时，白云机场几天下来依然有4名旅客滞留广州。为保
Fixed VAE	好航班延误情况下的航班保障工作，白云机场客运部昨日正式启动航
C-Enc	好航班延误情况下的航班保障工作，白云机场客运部昨日正式启动航
C+S-Enc	好航班延误情况下的航班保障工作，白云机场客运部昨日正式启动航
GT	好航班延误情况下的航班保障工作，白云机场客运部昨日正式启动航
Template: to keep an appointment with Heather.	
Style Ref	What seemed an undue period of
Fixed VAE	to keep an appointment with Heather.
C-Enc	to keep an appointment with Heather.
C+S-Enc	to keep an appointment with Heather.
GT	to keep an appointment with Heather.
at an intersection for the light to change. At this	
	A superfluous precaution for there was no
	at an intersection for the light to change. At this
	at an intersection for the light to change. At this
	at an intersection for the light to change. At this
	at an intersection for the light to change. At this

Figure 18: Handwritten Chinese and English text-line generation results under different encoder configurations. Red rectangles indicate cases where incorrect character structures are produced.

Table 10 shows that C+S-Enc achieves higher CR and AR scores than the frozen-VAE-based setup. This is because although C+S-Enc often generates characters with multiple extra strokes, these artifacts rarely mislead the downstream text-line recognition model, resulting in higher recognition metrics despite localized structural errors.

Compared with Chinese, English exhibits fewer character-level errors and generally acceptable styles, although the overall faint-stroke phenomenon still exists. This further indicates that Chinese handwriting is more challenging to model than English. Similarly, the C+S-Enc configuration yields more neatly shaped English words.

Overall, for both Chinese and English, the C-Enc and C+S-Enc settings underperform the frozen-VAE baseline. This may be partly attributed to the inherent alignment between the pretrained VAE’s latent space and the subsequent Diffusion Transformer, which provides a stronger and more stable representation than training the encoders from scratch.

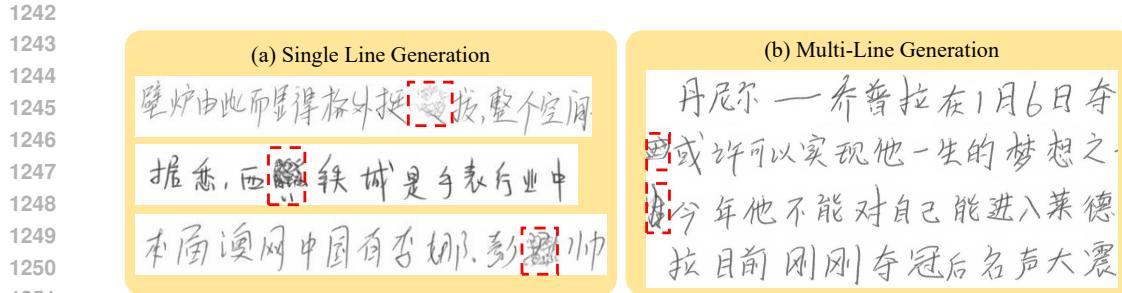


Figure 19: Visualization results of failure-case analysis. The red boxes highlight the erased marks mistakenly generated by the model.

A.5.6 ANALYSIS OF FAILURE CASES

A common failure mode we observed is that the model sometimes fills in the spaces of a text line with artifacts that resemble crossed-out or erased characters. This behavior stems from the training data: some authors habitually strike out a mistaken character and then write the correction beside it. When the generation is strictly conditioned by the content template, the correct action would be to preserve an empty region (i.e., a gap) for the removed character; instead, the model often produces residual “erasure” marks in that region. The problem is particularly pronounced at paragraph-level generation and near image boundaries. Representative examples are shown in Figure 19.

A.5.7 ANALYSIS OF STYLE–CONTENT CONTROLLABILITY AND DISENTANGLEMENT

We conduct an explicit style–content disentanglement experiment to verify that InkSpire maintains independent controllability over style and content, despite embedding style, content, and noise into a unified latent space. Given a fixed content template, we generate handwriting samples conditioned on multiple distinct style-reference lines sourced from different writers. This setup evaluates whether the model can vary stylistic attributes (e.g., stroke width, slant angle, character curvature) while keeping the textual content unchanged.

As shown in the visualization results (Figures 20):

- Pink boxes highlight the handwritten style of the letter *g*, and purple boxes highlight the style of the character sequence “th”. We observe that within each sample (e.g., sample a or b), all occurrences of *g* and *th* exhibit highly consistent writing patterns that closely match their respective style-reference samples. Meanwhile, samples a and b show strong stylistic divergence from each other, reflecting the differences in their reference styles.
- The red arrows indicate the slant direction of characters in samples c and d. Sample c shows a uniformly left-leaning slant across almost all words, matching its reference style, while sample d shows a consistently right-leaning slant at nearly uniform angles—again mirroring its style reference.

These observations confirm that the unified latent representation does not collapse style and content. Instead, InkSpire successfully preserves the content while expressing distinct stylistic attributes, demonstrating robust and independent controllability over both factors.

A.6 MORE VISUALIZATION EXAMPLES FOR MULTI-LINE TEXT GENERATION AND EDITING

We provide additional visualization results of InkSpire for multi-line handwritten text generation and character-level handwritten text editing. Figure 21 presents examples of English multi-line text generation, while Figure 22 shows the corresponding results for Chinese. Moreover, Figure 23 illustrates character-level editing results on English handwritten passages, and Figure 24 displays the editing results on Chinese passages.

1296
1297
1298
1299
1300
1301

1302
1303
1304
1305
1306
1307
1308
1309
1310
1311

1312
1313

1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327

Content Condition

than three thousand three fifty at the most. Now you could get quite a nice little semi-detached house in Grimstead for three thousand, that's where I live, just before the green belt, lovely and modern, you know. I forgot to tell you, we don't usually lend any house built earlier than 1918 "

Style Reference

"Good-oh! One for the gilt-room, darling, and

Generation

than three thousand three fifty at the most. Now you could get quite a nice little semi-detached house in Grimstead for three thousand, that's where I live, just before the green belt, lovely and modern, you know. I forgot to tell you, we don't usually lend any house built earlier than 1918 "

(a)

Style Reference

"And then we'll go for a stroll along the

Generation

than three thousand three fifty at the most. Now you could get quite a nice little semi-detached house in Grimstead for three thousand, that's where I live, just before the green belt, lovely and modern, you know. I forgot to tell you, we don't usually lend any house built earlier than 1918 "

(b)

Style Reference

"But I want it all protected. It's my ring!"

Generation

than three thousand three fifty at the most. Now you could get quite a nice little semi-detached house in Grimstead for three thousand, that's where I live, just before the green belt, lovely and modern, you know. I forgot to tell you, we don't usually lend any house built earlier than 1918 "

(c)

Style Reference

Hello, Cecil. Had a busy day? His mother said

Generation

than three thousand three fifty at the most. Now you could get quite a nice little semi-detached house in Grimstead for three thousand, that's where I live, just before the green belt, lovely and modern, you know. I forgot to tell you, we don't usually lend any house built earlier than 1918 "

(d)

Figure 20: Visualization results of Style–Content Controllability and Disentanglement. Pink boxes mark stylistic variations of “g”, purple boxes highlight “th”, and red arrows indicate the global slant direction.

1343
1344
1345
1346
1347
1348
1349

1350
1351
1352
1353
1354
1355**Ground Truth G**

"Good-oh! Give for the grill-room, darling, and tell George to bring in the *carte du jour*."
Bawley was a man of his word. He kept away from Service matters, was an excellent host and a splendid raconteur. The evening

Style Text:
"Good-oh! One for the grill-room, darling, and

"But I want it all protected. It's my ring."
"You just gave it back to me?" I didn't! I was joking. You know I was joking, Nigel." You shouldn't joke about serious things like engagements."

Style Text:
"But I want it!" she protested. "It's my ring."

Hello, Cecil. Had a busy day? His mother came into the hall as he opened the front door. He nodded irritably and, turning his back to her, contrived to slide the briefcase into ~~the~~ hiding between the do-it-yourself cupboard and the polished brass fourteen-pounder shell-case

Style Text:
"HELLO, CECIL. HAD a busy day?" His mother came into

"And then we'll go for a stroll along the beach. A moon like this mustn't be wasted." He gestured towards the bay where the full moon was just rising over the rocks, then laid his hand on hers. Gay smiled at him.

Style Text:
"And then we'll go for a stroll along the

'Good-bye, Sandra,' he said with a deadly finality. Watching him go, unable to speak, she felt that part of her was leaving him. She couldn't hate him...

Style Text:
"Good-bye, Sandra,' he said

Predict P

"Good-oh! Give for the grill-room, darling, and tell George to bring in the *carte du jour*."
Bawley was a man of his word. He kept away from Service matters, was an excellent host and a splendid raconteur. The evening

Target Text:
"Good-oh! tell George to bring in the *carte du jour*."
Bawley was a man of his word. He kept away from Service matters, was an excellent host and a splendid raconteur. The evening

"But I want it all protected. It's my ring."
"You just gave it back to me?" I didn't! I was joking. You know I was joking, Nigel." You shouldn't joke about serious things like engagements."

Target Text:
"But I want it all protected. It's my ring."
"You just gave it back to me?" I didn't! I was joking. You know I was joking, Nigel." You shouldn't joke about serious things like engagements."

Hello, Cecil. Had a busy day? His mother came into the hall as he opened the front door. He nodded irritably and, turning his back to her, contrived to slide the briefcase into ~~the~~ hiding between the do-it-yourself cupboard and the polished brass fourteen-pounder shell-case

Target Text:
"Hello, Cecil. HAD a busy day?" His mother came into

"And then we'll go for a stroll along the beach. A moon like this mustn't be wasted." He gestured towards the bay where the full moon was just rising over the rocks, then laid his hand on hers. Gay smiled at him.

Target Text:
"And then we'll go for a stroll along the beach. A moon like this mustn't be wasted."
He gestured towards the bay where the full moon was just rising over the rocks, then laid his hand on hers. Gay smiled at him.

'Good-bye, Sandra,' he said with a deadly finality. Watching him go, unable to speak, she felt that part of her was leaving him. She couldn't hate him...

Target Text:
"Good-bye, Sandra,' he said with a deadly finality. Watching him go, unable to speak, she felt that part of her was leaving him. She couldn't hate him...

1399
1400
1401
1402
1403

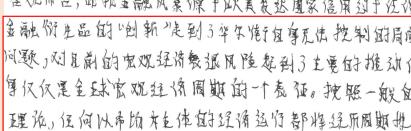
Figure 21: More examples of multi-line English paragraph generation.

1404
1405
1406
1407
1408
1409
1410
1411
1412

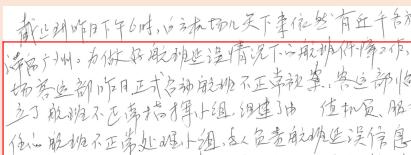
Ground Truth *G*

客观而言,此次金融风暴源于欧美发达国家信用过于泛滥金融衍生品的“创新”走到了华尔街自身无法控制的局面问题,对目前的宏观经济衰退风险起到了主要的推动作用仅仅是全球经济宏观经济发展周期的一个表征。按照一般的理论,任何以市场为主体的经济运行都将经历周期性的

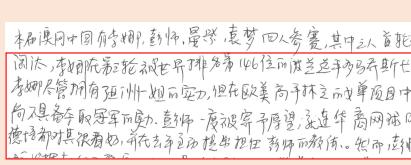
Style Text:
客观而言,本轮金融风暴源于欧美发达国家信用过于泛滥


截止到昨日下午6时,白云机场几天下来依然有近千名旅客滞留广州。为做好航班延误情况下的航班保障工作,白云机场客运部昨日正式启动航班不正常预案。客运部将成立了航班不正常指挥小组,组建了油量值机员、服务主任航班不正常处理小组,专人负责航班延误信息,并

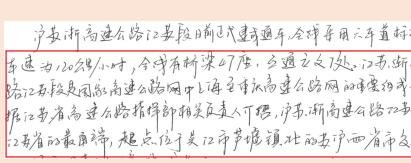
Style Text:
截止到昨日下午6时,白云机场几天下来依然有近千名


本屆澳網中國有李娜、彭帥、晏紫、袁夢四人參賽,其中三人首輪便淘汰,李娜在第二輪被世界排名第146位的波士頓選手瓦伊切拉法齊娜尽管拥有亚洲一流的实力,但在欧美高手林立的女单项目中,尚不具备夺取冠军的实力。彭帅一度被寄予厚望,就连华裔网球明星德约科维奇都对其很看好,并在去年主动提出担任彭帅的教练。然而,彭帅

Style Text:
本届澳网中国有李娜、彭帅、晏紫、袁梦四人参赛,其中三人首轮便


沪苏浙高速公路江苏段日前正式通车,今早采用六车道标准
车速为120公里/小时,今早有桥梁47座,互通立交7处。沪苏浙高速公路是国家高速公路网中上海至重庆高速公路网的重要组成部分,连接江苏省高速公路指挥部相关负责人介绍,沪苏浙高速公路江苏段起点位于吴江市芦墟镇北的苏沪两省市交界,江苏段的最南端,起点位于吴江市芦墟镇北的苏沪两省市交界。

Style Text:
沪苏浙高速公路江苏段日前正式通车,全线采用六车道标准


联邦公开市场委员会(FOMC)表示:“虽然短期融资市
压力稍微得到减轻,但金融市场情况却继续恶化;对于一
业以及个人来说,贷款变得更为困难。而且,新的信息反映
市场更加低迷,而劳工市场也出现疲软迹象。”其发布的
还称:“可预见的增长减缓的风险继续存在,美联储将

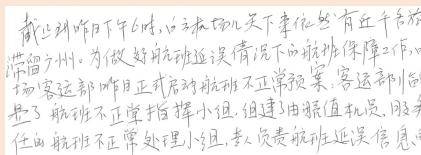
Style Text:
联邦公开市场委员会(FOMC)表示:“虽然短期融资市

1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457

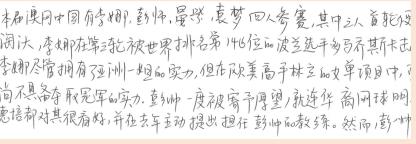
Predict *P*

客观而言,此次金融风暴源于欧美发达国家信用过于泛滥金融衍生品的“创新”走到了华尔街自身无法控制的局面问题,对目前的宏观经济衰退风险起到了主要的推动作用仅仅是全球经济宏观经济发展周期的一个表征。按照一般的理论,任何以市场为主体的经济运行都将经历周期性的

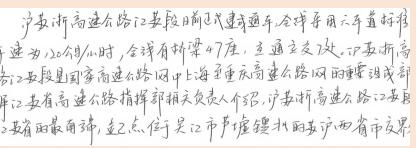
Target Text:
金融衍生品的“创新”走到了华尔街自身无法控制的局面问题,对目前的宏观经济衰退风险起到了主要的推动作用仅仅是全球经济宏观经济发展周期的一个表征。按照一般的理论,任何以市场为主体的经济运行都将经历周期性的


截止到昨日下午6时,白云机场几天下来依然有近千名旅客滞留广州。为做好航班延误情况下的航班保障工作,白云机场客运部昨日正式启动航班不正常预案。客运部成立了航班不正常指挥小组,组建了油量值机员、服务主任航班不正常处理小组,专人负责航班延误信息。

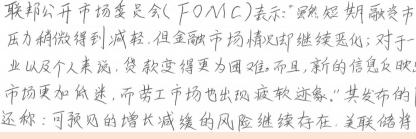
Target Text:
滞留广州。为做好航班延误情况下的航班保障工作,白云
机场客运部昨日正式启动航班不正常预案。客运部临时
成立了航班不正常指挥小组,组建了油量值机员、服务
主任航班不正常处理小组,专人负责航班延误信息。


本届澳网中国有李娜、彭帅、晏紫、袁梦四人参赛,其中三人首轮便淘汰,李娜在第二轮被世界排名第146位的波士顿选手瓦伊切拉法齐娜淘汰,尽管拥有亚洲一流的实力,但在欧美高手林立的女单项目中,尚不具备夺取冠军的实力。彭帅一度被寄予厚望,就连华裔网球明星德约科维奇都对其很看好,并在去年主动提出担任彭帅的教练。然而,彭帅

Target Text:
淘汰,李娜在第三轮被世界排名第146位的波兰选手多马乔斯卡击
李娜尽管拥有亚洲一流的实力,但在欧美高手林立的女单项目中,李
尚不具备夺取冠军的实力。彭帅一度被寄予厚望,就连华裔网球明星
德约科维奇都对其很看好,并在去年主动提出担任彭帅的教练。然而,彭帅


沪苏浙高速公路江苏段日前正式通车,今早采用六车道标准
车速为120公里/小时,今早有桥梁47座,互通立交7处。沪苏浙高
速公路是国家高速公路网中上海至重庆高速公路网的重要组成部分,连接江苏省高速公路指挥部相关负责人介绍,沪苏浙高速公路江苏段起点位于吴江市芦墟镇北的苏沪两省市交界,江苏段的最南端,起点位于吴江市芦墟镇北的苏沪两省市交界。

Target Text:
车速为120公里/小时,今早有桥梁47座,互通立交7处。沪苏浙高
速公路是国家高速公路网中上海至重庆高速公路网的重要组成部分,连接江苏省高速公路指挥部相关负责人介绍,沪苏浙高速公路江苏段起点位于吴江市芦墟镇北的苏沪两省市交界。


联邦公开市场委员会(FOMC)表示:“虽然短期融资市
压力稍微得到减轻,但金融市场情况却继续恶化;对于一
业以及个人来说,贷款变得更为困难。而且,新的信息反映
市场更加低迷,而劳工市场也出现疲软迹象。”其发布的
还称:“可预见的增长减缓的风险继续存在,美联储将

Target Text:
压力稍微得到减轻,但金融市场情况却继续恶化;对于一
业以及个人来说,贷款变得更为困难。而且,新的信息反映
市场更加低迷,而劳工市场也出现疲软迹象。”其发布的
还称:“可预见的增长减缓的风险继续存在,美联储将

1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468

Original *G*

"Good-oh! **One** for the grill-room, darling, and tell George to bring **in** the **carte du jour**."
Bawley was a man of his word. He **kept** away from Service **matters**; was an excellent host and a splendid raconteur. The **evening**:

Target Text:
"Good-oh! **One** for the grill-room, darling, and tell George to bring **in** the **carte du jour**."
Bawley was a man of his word. He **kept** away from Service **matters**, was an excellent host and a splendid raconteur. The **evening**

Kello, Cecil. Had a **busy** day? His mother **came** into the **hall** as he opened the **front** door. He nodded irritably and, turning his back to her, contrived to slide the briefcase into **hiding** between the do-it-yourself cupboard and the polished **brass** fourteen-pounder shell-case

Target Text:
"HELLO, CECIL. HAD a **busy** day?" His mother came into the **hall** as he opened the **front** door. He nodded irritably and, turning his back to her, contrived to slide the briefcase into **hiding** between the do-it-yourself cupboard and the polished **brass** fourteen-pounder shell-case

"But I want it **safe** protected. It's my ring."
"You just **gave** it back to **me**!" I didn't! I was joking. You **knew** I was joking, **Nigel**! "You shouldn't joke **about** serious things like engagements!"

Target Text:
"But I want it **safe** protected. It's my ring."
"You just **gave** it back to **me**!" "I didn't! I was joking. You **knew** I was joking, **Nigel**." "You shouldn't joke **about** serious things like engagements."

1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511

Edited *P*

"Good-oh! **Two** for the grill-room, darling, and tell George to bring **on** the **carte du jour**."
Bawley was a man of his word. He **kept** away from Service **reasons**; was an excellent host and a splendid raconteur. The **morning**:

Target Text:
"Good-oh! **Two** for the grill-room, darling, and tell George to bring **on** the **carte du jour**."
Bawley was a man of his word. He **keep** away from Service **reasons**, was an excellent host and a splendid raconteur. The **morning**

Kello, Cecil. Had a **busy** day? His mother **came** into the **hall** as he opened the **back** door. He nodded irritably and, turning his back to her, contrived to slide the briefcase into **hiding** between the do-it-yourself cupboard and the polished **brass** fourteen-pounder shell-case

Target Text:
"HELLO, CECIL. HAD a **busy** day?" His mother came into the **hall** as he opened the **back** door. He nodded irritably and, turning his back to her, contrived to slide the briefcase into **hiding** between the do-it-yourself cupboard and the polished **brass** fourteen-pounder shell-case

"But I want it **safe** protected. It's my ring."
"You just **gave** it back to **us**!" I didn't! I was joking. You **knew** I was joking, **Nolan**! "You shouldn't joke **around** serious things like engagements!"

Target Text:
"But I want it **safe** protected. It's my ring."
"You just **give** it back to **us**!" "I didn't! I was joking. You **knew** I was joking, **Nolan**." "You shouldn't joke **around** serious things like engagements."

Figure 23: More examples of English paragraph editing.

1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524

Original *G*

客观而言，此轮金融风暴源于欧美发达国家信用过于泛滥金融衍生品的“创新”走到华尔街自身无法控制的局面问题，对目前的宏观经济衰退风险起到了主要的推动作用。仅仅是全球宏观经济增长周期的一个表现，按照一般的理论，任何以市场为主体的经济运行都将经历周期性的

Target Text:

客观而言，此轮金融风暴源于欧美发达国家信用过于泛滥金融衍生品的“创新”走到了华尔街自身无法控制的局面问题，对目前的宏观经济衰退风险起到了主要的推动作用。仅仅是全球宏观经济增长周期的一个表现。按照一般的理论，任何以市场为主体的经济运行都将经历周期性的

本届澳网中国有李娜、彭帅、晏紫、袁梦四人参赛，其中三人首场便淘汰，李娜在第三轮被世界排名第146位的波兰选手多马乔斯卡击败。李娜尽管拥有亚洲一姐的实力，但在中美高手林立的女单项目中，尚不具备夺取冠军的实力。彭帅一度被寄予厚望，就连华裔网球明星德普都对其很看好，可从去年年底她出任彭帅的教练。然而，彭帅

Target Text:

本届澳网中国有李娜、彭帅、晏紫、袁梦四人参赛，其中三人首场便淘汰，李娜在第三轮被世界排名第146位的波兰选手多马乔斯卡击败。李娜尽管拥有亚洲一姐的实力，但在中美高手林立的女单项目中，尚不具备夺取冠军的实力。彭帅一度被寄予厚望，就连华裔网球明星德普都对其很看好，可从去年年底她出任彭帅的教练。然而，彭帅

联邦公开市场委员会(FOMC)表示：“虽然短期融资市场的压力稍微得到减轻，但金融市场情况却继续恶化；对于一业以及个人来说，贷款变得更为困难。而且，新的信息反映：市场更加低迷，而劳工市场也出现疲软迹象。”其发布的还称：可预见的增长减缓的风险继续存在，美联储将

Target Text:

联邦公开市场委员会(FOMC)表示：“虽然短期融资市场的压力稍微得到减轻，但金融市场情况却继续恶化；对于一业以及个人来说，贷款变得更为困难。而且，新的信息反映：市场更加低迷，而劳工市场也出现疲软迹象。”其发布的还称：可预见的增长减缓的风险继续存在，美联储将

Edited *P*

客观而言，此轮金融风暴源于欧美发达国家信用过于泛滥金融衍生品的“创新”走到了华尔街自身无法控制的局面问题，对目前的宏观经济衰退风险起到了主要的推动作用。仅仅是全球宏观经济增长周期的一个表现，按照一般的理论，任何以市场为主体的经济运行都将经历周期性的

Target Text:

客观而言，此轮金融风暴源于欧美发达国家信用过于泛滥金融衍生品的“创新”走到了华尔街自身无法控制的局面问题，对目前的宏观经济衰退风险起到了主要的推动作用。仅仅是全球宏观经济增长周期的一个表现。按照一般的理论，任何以市场为主体的经济运行都将经历周期性的

本届澳网中国有李娜、彭帅、晏紫、袁梦四人参赛，其中三人首场便淘汰，李娜在第三轮被世界排名第146位的波兰选手多马乔斯卡击败。李娜尽管拥有亚洲一姐的实力，但在中美高手林立的女单项目中，尚不具备夺取冠军的实力。彭帅一度被寄予厚望，就连华裔网球明星德普都对其很看好，可从去年年底她出任彭帅的教练。然而，彭帅

Target Text:

本届澳网中国有李娜、彭帅、晏紫、袁梦四人参赛，其中三人首场便淘汰，李娜在第三轮被世界排名第146位的波兰选手多马乔斯卡击败。李娜尽管拥有亚洲一哥的实力，但在中美高手林立的女单项目中，尚不具备夺取冠军的实力。彭帅一度被寄予厚望，就连华裔网球明星德普都对其很看好，并在去年主动提出担任彭帅的教练。然而，彭帅

联邦公开市场委员会(FOMC)表示：“虽然长期融资市场的压力稍微得到减轻，但金融市场情况却继续恶化；对于一业以及个人来说，贷款变得更为困难。而且，旧的信息反映：市场更加低迷，而劳工市场也出现疲软迹象。”其发布的还称：可预见的增长增速的风险继续存在，美联储将

Target Text:

联邦公开市场委员会(FOMC)表示：“虽然长期融资市场的压力稍微得到减轻，但金融市场情况却继续恶化；对于一业以及个人来说，贷款变得更为困难。而且，旧的信息反映：市场更加低迷，而劳工市场也出现疲软迹象。”其发布的还称：可预见的增长增速的风险继续存在，美联储将

1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565

Figure 24: More examples of Chinese paragraph editing.